Journal of Economic Literature
Vol. XLIII (March 2005), pp. 65–107

Economic Theory and Experimental Economics

Larry Samuelson*

1. *Introduction*

Game theory had its beginnings in economics as a separate topic of analysis, practiced by a cadre of specialists. It has since become commonplace. Every economist is acquainted with the basic ideas, often without notice, and there is free movement between the use of game theory and other techniques. This incorporation as a standard economic tool has helped shape the nature of game theory itself—the mix of questions has changed and more attention has been devoted to how game theoretic models are to be interpreted as capturing economic interactions.

Mathematical economics and econometrics have each similarly progressed from being a topic pursued by a band of specialists to becoming a sufficiently familiar tool as to be used without comment. In the process, each has been shaped by issues arising in economic applications.

Experimental economics is currently making its transition from topic to tool.[1] Once viewed skeptically by many economists, experiments have become commonplace. Once again, this transition has involved changes both in the way economists view experimental methods and in the experimental methods themselves.

This paper explores one aspect of this integration of experimental economics into economics. How can we usefully combine work in economic theory and experimental economics? What do economic theory and experimental economics have to contribute to one another, and how can we shape their interaction to enhance these contributions?

There is already plenty of work that insightfully integrates theory and

* Samuelson: University of Wisconsin. I thank Jim Andreoni, Jakob K. Goeree, Roger Gordon, John McMillan, Georg Nöldeke, and three referees for helpful comments. I thank the Economic Science Association for the invitation to give a talk at the 2003 Pittsburgh meetings that developed into this paper. I thank the National Science Foundation (SES-0241506) and Russell Sage Foundation (82-02-04) for financial support.

[1] For example, the *Journal of Economic Literature's* "Mathematical and Quantitative Methods" classification section includes a "Design of Experiments" subsection, and a Nobel prize has been given for experimental work. At the same time, training in experimental methods has not yet joined basic econometrics or game theory as a standard part of the first-year graduate curriculum. Alvin E. Roth (1993) provides a history of early work in experimental economics. Roth (1995) continues this history and provides a more detailed discussion of recent experimental work. Roth (1991) proceeds further with some thoughts on the future of experiments in economics.

experiments.[2] However, the methods for putting the two together are still developing. The goal here is to examine the issues involved in this development. Much can be gained by combining economic theory and experiments, but doing so calls for thinking carefully about the way we do theory as well as experiments.

2. *An Example*

It is helpful to begin with an example in which experimental results and economic theory have constructively mingled. This example illustrates the ideas that will be developed more generally in section 3, illustrated in section 4, and then extended in section 5.

In 1965, Reinhard Selten (1965) introduced the concept of a subgame-perfect equilibrium. Subgame perfection is now taken for granted, in the sense that a paper whose conclusion hinged upon an equilibrium that was *not* subgame perfect would have a great deal of explaining to do.

Some years later, Werner Güth, Rolf Schmittberger, and Bernd Schwarze (1982) performed a simple experiment, examining what has come to be known as the *ultimatum game*. Player 1 makes a proposal for how a sum of money is to be split between players 1 and 2. Player 2 then either accepts, implementing the proposal, or rejects, in which case the interaction ends with zero payoffs for each. This is the type of game—perfect information, two players, only one move per player—in which subgame perfection is often viewed as being obviously compelling. In any subgame-perfect equilibrium of the ultimatum game, player 1 makes and player 2 accepts a proposal that gives player 2 at most one penny (or one of whatever is the smallest monetary unit available). In contrast, Güth, Schmittberger, and Schwarze obtained results that have been echoed by an ever-growing list of subsequent studies. The modal proposal is typically to split the sum of money evenly. If player 1 asks for two-thirds or more of the surplus, he stands a good chance of being rejected.

We thus have a marked contrast between theory and experiment. A common initial reaction was to dismiss the laboratory environment as uninteresting. Why should we be interested in how experimental subjects play an artificial game for token amounts of money? Borrowing a term from experimental psychology, this is a question of external validity: is the experimental environment sufficiently close to the situation of interest to be informative? In this case, for example, is the laboratory environment close enough to the situations envisaged by contract theorists when they assume that subgame-perfect equilibria appear in the ultimatum games embedded in their models?

One way of gaining some perspective on such questions is to turn them around. How special is the laboratory environment generating the experimental results? Can we link the results to aspects of the experimental environment that appear to be especially artificial, or do they appear to be robust? In the case of the ultimatum game, a long string of experiments has investigated the effects of playing for larger amounts of money, playing in different countries and cultures, playing with differing degrees of anonymity, playing with different amounts of experience, playing games of different length, and playing with different types of opponents.[3] Some of these

[2] Vincent P. Crawford (1997) and Roth (1988) explore the interaction between economic theory and experiments, each arguing (as does this paper) that there are good reasons for thinking about experiments when doing economic theory as well as thinking about theory when doing experiments.

[3] For example, Lisa A. Cameron (1999), Elizabeth Hoffman, Kevin A. McCabe and Vernon L. Smith (1996), and Robert L. Slonim and Roth (1998) (larger payoffs); Joseph Henrich (2000), Henrich, Robert Boyd, Samuel Bowles, Colin F. Camerer, Ernst Fehr, Herbert Gintis, and Richard McElreath (2001), and Roth, Vesna Prasnikar, Masahiro Okuno-Fujiwara, and Shmuel Zamir (1991) (different countries and cultures); Gary E. Bolton and Rami Zwick (1995), Hoffman, McCabe, and Smith (1996), and Hoffman, McCabe, Shachet, and Smith (1994) (anonymity); David Cooper, Nick Feltovich, Roth, and Zwick (2002) (experience); Robert Forsythe, Joel L. Horowitz, N. E. Savin, and Martin Sefton (1995) and Glenn W. Harrison and McCabe (1992) (length); and Sally Blount (1995), Harrison and McCabe (1996), and Eyal Winter and Zamir (1997) (opponents).

variations matter, and there is much to be learned about which matter more than others. However, no combination of conditions has been found that produces the subgame-perfect equilibrium outcome sufficiently reliably as to allow us to dismiss the remaining experimental results. The mounting evidence suggests that the ultimatum game has something to tell us about behavior. One can often find reasons to dismiss any single experiment, but cannot ignore such a large and varied body of work.

Attention then turns to the theory. What implications for economic theory do the experimental results have? Perhaps none. We know that any theory is a deliberate approximation, and hence that there must be *some* circumstances under which it fails. Could it be that the theory is meant for settings not captured by the experiments, and that the theory is still useful in the applications for which it is intended? In this spirit, Ken Binmore, Avner Shaked, and John Sutton (1985) argue that the ultimatum game features an atypically asymmetric division of bargaining power, making subgame perfection unrealistically demanding, and that models built around subgame perfection might be a better match for two-stage bargaining games that feature a less extreme (though still asymmetric) distribution of bargaining power. Their experiments produced outcomes much closer to the subgame-perfect equilibrium in two-stage bargaining games. Are we then to assume that the subgame-perfect equilibrium is a useful model of behavior in bargaining models, as long as we stay away from models with equilibria that are too asymmetric? And if so, what does "too asymmetric" mean?

Once again, we can seek insight in the ensuing body of experimental work. Subsequent experiments have examined bargaining games with varying degrees of asymmetry in bargaining power (see Douglas D. Davis and Charles A. Holt [1993] for a summary). Departures from equilibrium are often much less pronounced than in the ultimatum game, but the data still does not invariably reflect equilibrium play. However, we do not expect theories to make *exact* predictions. How close is close enough? When are experimental results within the margin of approximation that is inevitably built into a theory, and when do they indicate that the theory is on the wrong track? There is typically no obvious standard for answering these questions. One can then imagine Davis and Holt's summary figure (Figure 5.6) being regarded as evidence for both the success and the failure of conventional bargaining models, depending upon one's point of view.

A return to the theory is again helpful, this time with an eye toward finding within the theory some guide for evaluating the experimental results. Glen W. Harrison (1989, 1992) (see also Robert Drago and John S. Heywood [1989]) suggests one approach. A cornerstone of the relevant theory is that people maximize their expected payoffs. In light of this, a natural measure for evaluating the theory is the payoff losses subjects incur as a result of not behaving as predicted. The larger are these losses, the stronger is the evidence that the theory has missed something. Harrison argues that in the case of auctions, seemingly large departures from equilibrium behavior often translate into very small payoff losses, suggesting that the contrast between theory and behavior is not nearly as large as it first appears. Drew Fudenberg and David K. Levine (1997) turn a similar eye toward a variety of other games, including the ultimatum game. They find that behavior in many of these games is consistent with subjects' holding beliefs about others' behavior that is consistent with their experience and against which they suffer relatively small payoff losses. This again suggests that the theory may capture important elements of behavior, despite seemingly unencouraging experimental results. At the same time, however, payoff losses in the ultimatum game are relatively large compared to many other experiments. Rejecting an offer often involves a significant sacrifice, regardless of what one believes about how others are playing. It is

then harder to argue here that one can rationalize nonequilibrium behavior simply by arguing that the players are nonetheless achieving approximately equilibrium payoffs.

Binmore, John Gale, and Larry Samuelson (1995) and Alvin E. Roth and Ido Erev (1995) suggest an alternative approach to assessing how close observed behavior is to the predictions of the theory. Why should we expect equilibrium behavior in the first place? The traditional answer in economics is not that equilibria spring to life as a result of sheer calculation or external organization, but rather that behavior is pushed toward equilibrium by an adjustment or learning process that continually puts pressure on players to alter *nonequilibrium* behavior. Adopting this view, how strong are the incentives for players to adjust nonequilibrium behavior in simple bargaining games?[4] The stronger are these incentives, the stronger is the experimental evidence that the theory has missed something. In the case of bargaining games, it turns out that these incentives can be quite weak. Even small amounts of noise or imperfection can cause the learning process to get stuck, for long or even indefinite periods of time, far away from a neighborhood of the subgame-perfect equilibrium. We thus again have a suggestion that the observed behavior may not be too far from equilibrium, by at least one measure of "too far." Motivated by similar considerations, a literature on learning and its relationship to experimental behavior has developed.[5] However, two difficulties now arise. First, what can the subjects, especially responders, possibly have to learn in a game so simple as the ultimatum game? Without a clear answer to this question, learning models are difficult to interpret. We return to this question in section 5. Second, learning theories have proven to be cumbersome tools with which to examine strategic interactions. A successful theory trades off its explanatory power with its ease of use. It has not been easy to formulate learning models that rival equilibrium theories in terms of readily yielding sharp predictions.[6]

Perhaps one should view the connection between theory and experiment differently. Instead of asking whether the theory gets the behavior right, and then wrangling over how the distance between experimental and theoretical outcomes is to be measured and interpreted, let us ask whether the theory captures the important considerations shaping the behavior. This directs attention away from the point predictions of the theory and toward its comparative statics. For example, experimental behavior that consistently responds to changes in discount rates as predicted by the theory of bargaining might lead us to believe that the theory has identified an important role for impatience in shaping behavior, even if the theory is not complete enough to capture every aspect of behavior. This emphasis on comparative statics pushes experimental analysis closer to methods familiar in other areas of economics.

The results in this respect for bargaining theory are mixed. For example, Binmore, Peter Morgan, Shaked, and Sutton (1991) and Binmore, Shaked, and Sutton (1989) report experiments in which behavior responds to the difference between a voluntarily exercised and involuntarily exercised outside option in a direction consistent with theoretical predictions. However, Jack Ochs

[4] In spirit, this is close to asking how far realized payoffs fall short of the payoffs that could be obtained by playing a best response. The difference is that the incentives for adjusting one's behavior are now taken to be not the payoffs promised by perfect optimization, but rather the incentives to pursue the potentially imperfect learning process that shapes behavior.

[5] See Camerer (2003) and Drew Fudenberg and David K. Levine (1998) for examples. Raymond Battalio, Samuelson, and John Van Huyck (2001) report an experiment linking the speed of learning and the incentives for adjusting one's strategy, providing some hint that learning can be important.

[6] Ed Hopkins (2002) provides an indication of why it can be difficult to identify the learning process behind experimental behavior.

and Roth (1989) report an experiment in which behavior does not respond to the discount factor and the length of the game consistently with the predictions of subgame perfection. This latter finding is all the more disconcerting because the role of impatience is viewed as one of the key insights of noncooperative bargaining models.[7] These results suggest that rates of impatience may be less central, and the prospect of a breakdown in negotiations more important, than captured by the original models.

Taken together, the body of experimental evidence suggests that our simplest theories of bargaining leave some aspects of behavior unexplained. This is interesting, but is most useful if the experiments also suggest how we might construct a more encompassing account of behavior. This brings us to the question, again borrowed from experimental psychology, of internal validity. How do we assess whether our interpretation of an experimental result captures the relevant aspects of the experimental situation and the resulting behavior, and hence points the way to a better understanding of the behavior and to better theoretical models of that behavior? For example, Güth, Schmittberger, and Schwarze (1982) (see also Güth and Reinhard Tietz [1982]) interpret their results as indicating that subjects' behavior is shaped primarily by considerations of fairness. If this is the case, then we may be on the road to a new "fairness theory" of behavior. We might work with familiar bargaining models, but with quite different views of how people behave in these models. Notice that this assessment differs markedly from the hints with which the previous paragraph concluded, under which we would retain the basic view of self-interested behavior but revise the structure of the model.

An appeal to fairness has an intuitive ring to it. It is hard to believe that fairness does not play a role in our lives, or that extremely asymmetric allocations would not strike one as unfair. It also seems quite natural that these considerations would carry over into behavior in bargaining experiments. Here, however, we return to a theme that appeared in connection with learning models and that runs throughout this essay. The relevant question for evaluating a theory is not so much whether it is "correct," but whether it can be readily and usefully applied to a sufficiently broad range of settings. The difficulty with appeals to fairness is that they too often have an "I know it when I see it" quality that makes them particularly cumbersome to use. Vesna Prasnikar and Roth (1992) develop this idea, reporting experimental results showing that, under some circumstances, experimental subjects do settle on extremely asymmetric allocations (see also James Andreoni, Paul M. Brown, and Lise Vesterlund [2002] and Harrison and Jack Hirshleifer [1989]). This appears to suggest that we have been too hasty in concluding that concerns for fairness routinely push people away from asymmetric allocations. However, the extreme allocations in Prasnikar and Roth's best-shot game Pareto dominate the less asymmetric allocations. In response, it is tempting to refine the notion of fairness, viewing it as inducing an antipathy to asymmetric allocations, but an antipathy that is tempered when asymmetric allocations have efficiency properties that symmetric allocations lack. Adding the best-shot experimental results to our portfolio may thus suggest that fairness is important after all, but is a more subtle notion than simply a concern for equality or symmetry.[8]

[7] The contrast between these two results becomes sharper in the context of section 4, which suggests that one should be especially disappointed when a theory fails to exhibit behavior integral to its original structure, such as the appropriate sensitivity to discount rates, but especially pleased when the theory successfully extends to originally novel questions, such as the effect of outside options.

[8] Prasnikar and Roth (1992) investigate these possibilities by examining a market game in which asymmetric equilibrium outcomes appear that do not Pareto dominate the symmetric outcomes.

There is much that is appealing about this argument, but it illustrates the difficulties of working with such an elusive concept as fairness. The more subjective or context-dependent is the idea of fairness, the less useful it becomes as a component of a theory, regardless of how important it is in shaping behavior.

Making progress in interpreting seemingly anomalous experimental results thus requires making the idea of fairness, or whatever else it is that one imagines affecting players' behavior, sufficiently precise. A first question is theoretical: can we do so with conventional theoretical techniques, or are we dealing with something quite different? Are we dealing with a world to which the underlying structure of economic models applies—people maximize, they balance competing objectives, they respond to variations in the constraints on how these objectives can be traded against one another—even if they are concerned with something other than simply how much money they make? Or are such models of behavior simply on the wrong track? Andreoni and John Miller (2002) provide some insight into this question through experiments in which a dictator faces a variety of exchange rates between the payoffs that the dictator can keep or allocate to a recipient (while Andreoni, Marco Castillo, and Ragan Petrie [2003] do much the same for the ultimatum game). As is often the case in such games, their results are not consistent with the proposition that all subjects care only about how much money they receive. However, their results are consistent with the claim that most subjects have stable preferences satisfying revealed-preference axioms. Whatever motivates the subjects, whether money or fairness or something else, it is something that we can model with the familiar optimization tools of economics, without abandoning rational behavior as a unifying principle.

The next task is again theoretical: fitting some more encompassing model of individual behavior into standard models of bargaining. Gary E. Bolton and Axel Ockenfels (2000) and Ernst Fehr and Klaus M. Schmidt (1999) (see also Bolton [1991]) offer models of preferences that capture a concern for fairness. Each is centered around a utility function that involves one's own payoff and the payoff of one's opponent, and that exhibits some aversion to payoff inequality.

One tempting reaction to these models is that nothing so simple could possibly capture the complexity of human behavior. Pursuing this view, it is not hard to find evidence that some factors are missing from these models.[9] However, such criticisms miss the point. It is again important to recall that one purpose of any theory is to judiciously choose considerations to neglect. The ability to find some circumstances in which the theory does not work perfectly is then not by itself cause to reject the theory. While they may still be incomplete, the models offered by Bolton and Ockenfels (2000) and Fehr and Schmidt (1999) have the key virtue that their predictions are clear and they can easily be extended to encompass novel situations. This allows us to confirm that these models predict behavior matching that of standard models in a wide variety of circumstances in which the latter appear to be applicable, to confirm that they capture the apparent fairness considerations that operate in the bargaining models that motivated their construction, and to investigate the extent to which they apply to new applications. This is just what we need to make progress, and is what economic theory must do if we are to effectively combine theory and experiments.

A variety of alternative and more elaborate models have appeared, many enriching the

[9] Among others, Ken Binmore, John McCarthy, Giovanni Ponti, Samuelson, and Avner Shaked (2002) and Armin Falk, Fehr, and Urs Fischbacher (2003) report experimental results indicating that preferences must depend upon more than simply payoffs, even the payoffs of all players.

theory by incorporating elements in preferences beyond simply the final allocation of payoffs.[10] There is considerable work to be done in assessing and synthesizing these models, work that will require a continual interplay between economic theory and experiments. How is this interplay to proceed? It will be helpful to develop some of the ideas raised in the course of this example more generally.

3. *A Theoretical Perspective*

This section opens a more general discussion with a theoretical perspective, in the form of a model of economic theory and economic experiments. The idea is to provide a precise way of talking about what a theory is, what an experiment is, and how the two might be related.

3.1 *A Model*

The environment. The model begins with the assumption that there is an objective environment or "real world" to be studied, represented by a function

$$F : X^{\infty} \to S^{\infty},$$

where X and S are finite sets and X^{∞} and S^{∞} are their infinite-dimensional products.[11] We think of the function F as taking in information, given by an element of the set X^{∞}, that defines a *situation* of interest. This information might define an extensive-form game, or a set of lotteries from which one is to choose, or a market or an economy. With each such situation, the function F associates an output from the set S^{∞}.[12] Depending on the situation, this output might be an equilibrium of the game, or a selected lottery, or a market price or a competitive equilibrium.

We can view each of the dimensions of X^{∞} and S^{∞} as corresponding to a property or characteristic that a situation or an output might have, with the sets X and S providing the language in which one describes such properties. The details of the sets X and S are not particularly important. What does matter is that there is a potentially endless list of relevant properties, sufficiently many that neither theoretical nor experimental work could ever hope to describe *every* aspect of reality. We ensure this in the model by assuming that there are infinitely many such properties, so that the sets of inputs and outputs are the infinite products X^{∞} and S^{∞}.[13]

We think of the environment as generating situations which are then transformed into outcomes by the function F. We let ρ denote a probability distribution describing the process that generates situations. We think of a theory or an experiment as being a tool for understanding the function F. In

[10] Bolton (1991) offers an early model in which preferences depend upon others' payoffs, while Matthew Rabin (1993), building on the theory of psychological games (John Geanakoplos, David Pearce, and Ennio Stacchetti [1989]), is an early example of how one might make the idea of fairness theoretically operational. Other examples include Gary Charness and Rabin (2002), Martin Dufwenberg and Georg Kirchsteiger (2004), Fehr and Simon Gächter (2000), Levine (1998), and Uzi Segal and Joel Sobel (2003). Andreoni and Samuelson (2003) report experimental results that explore, in a somewhat different setting, some of the key features of these models.

[11] This discussion thus avoids questions concerning the existence of an objective reality have been raised from widely differing perspectives. Physicists argue that quantum phenomena are not determined independently of attempts to measure them, while some social scientists argue that nothing objective exists independently of the observer, who constructs reality as she observes it. Such concerns can be relevant for economics. For example, could experimental procedures designed to elicit valuations affect those valuations? We return to this set of issues in section 5.1.

[12] Again, the model skirts a philosophical issue, concerning whether the world is deterministic or random. We adopt the technical convenience that outcomes are deterministic (conditional on being able to identify the situation completely), though in practice we can identify only finite approximations of situations, with outcomes that then appear to be random (conditional on this information). A random-world view requires only additional notation in order to accommodate an extra layer of probability distributions in the model.

[13] The sets X^{∞} and S^{∞} can be viewed as a short-hand for sets that are finite but prohibitively large. Daniel C. Dennett's *Library of Mendel* (1995) provides the setting for an intriguing discussion of large finite sets.

the absence of any constraints, of course, one would simply work with F itself. Unfortunately, the function F is too complicated to work with directly. The idea is then to combine theory and experimental work to produce tools that are simple enough to be used, while capturing enough of F to be useful.

Theory. Like the function F that describes the environment, a theory takes in information concerning a situation and provides information concerning the corresponding outcome. However, instead of taking in all of the information contained in an element of the set X^∞, we model the theory as making use only of the dimensions $1,\ldots,N$, for some finite N. Similarly, instead of specifying every detail of the output, the theory provides information only about the dimensions $1,\ldots, M$, for some finite M. Let X^N be the set of N-tuples corresponding to the first N dimensions of the set X^∞, and let S^M similarly be M-tuples whose elements correspond to the first M dimensions of S^∞. A *theory* is then a function

$$f : X^N \to \Delta\Delta S^M,$$

for some N and M, where ΔS^M is the set of probability measures over S^M and $\Delta\Delta S^M$ is the set of probability measures over ΔS^M.

The restriction to finite N and M captures the fact that a theory does not make use of all of the information defining a situation, nor does it specify every detail of the output.[14] Instead, one of the challenges in crafting a theory is to choose its inputs and outputs, i.e., to choose N and M, so as to include relevant information and neglect relatively unimportant details. For example, a theory about labor force participation rates may use information on wage rates, marital status, age and educational attainment, but may neglect information concerning foreign exchange rates. The theory's output may provide information about how much time an individual devotes to leisure, but may say nothing about which activities consume this time.

As a first approximation, we might think of the theory as choosing an output from S^M. However, given that the theory's input leaves some details of the situation unspecified, it is more natural to view the theory as producing a probability distribution over the outputs in S^M (i.e., an element of ΔS^M). We then interpret the random output as reflecting the uncontrolled realization of those aspects of the situation that are *not* captured by X^N. For example, a theory of labor force participation may provide an expected participation rate, as a function of an individual's age and education, but would view actual participation as being randomly distributed around this expected value, reflecting other, unobserved characteristics.[15]

It is useful to go one step further and allow the theory to produce an element of $\Delta\Delta S^M$, the space of distributions over distributions. We may have more information concerning the likely values and implications of some of the unmodeled features of a situation than of others. We may then have a distribution over the realizations of the features about which we have relatively good information, each in turn inducing a distribution over outcomes. For example, an analyst asked to predict the outcome of the next presidential election might begin with the question of whether the economy will then be healthy or in recession. The analyst's theory may involve a distribution over which of these is likely to be the case and, conditional on either, a distribution

[14] While it is intuitive that a theory cannot make use of all the information in the environment, there is in principle no reason why it should be restricted to the first N dimensions of X^∞. Why not a theory that makes use of the information in dimensions 1, 3, and 14 and ignores the rest? There is no loss in assuming that, whatever theory we have, the dimensions of X^∞ are arranged so that the theory makes use of an initial string of them.

[15] The idea that the theory produces a distribution over outcomes is perhaps most familiarly exploited by weather forecasters, who regularly announce probabilities of rain, but also appears routinely in economics. We return to this idea in section 4.1.2.

over likely outcomes of the election. Similarly, an economist asked to analyze the market for skilled labor might begin with a distribution over likely macroeconomic conditions, each of which in turn induces a distribution over conditions in the relevant labor market. In a model of labor force participation, one's education and age may induce a distribution over participation decisions that itself depends randomly on labor market conditions. Who is to say whether the probability that the weather forecaster attaches to rain is not itself chosen randomly?

Experiments. An experiment similarly associates an output with an input. The experiment again begins with an element of X^N (for some finite N), which we denote by x^N and refer to as the experimental design. This design fixes those features of the environment captured in the N dimensions of X^N. For example, the design x^N may specify how much money the subjects earn in various circumstances.

The actual input to the experiment is a situation, i.e. an element of X^∞ (denoted by x^∞), that matches x^N on the first N dimensions.[16] The idea here is that the experimental design fixes those details of the environment described by x^N, while leaving others uncontrolled. For example, the design may leave uncontrolled the wealth levels of the subjects.

Let F^M denote the function comprising the first M dimensions of F. Given an experimental design x^N and a corresponding input x^∞, the experiment consists of an observation of the form:

$$F^M(x^\infty)$$

for some M. Hence, an experiment consists of a partial description ($F^M(x^\infty)$) of the output of the function F that describes the environment, evaluated at one of a collection of possible inputs ($x^\infty \in X^\infty$) that share the features (x^N) determined by the experimental design.[17]

It may seem counterintuitive to characterize the experiment as yielding realizations of F, since experiments are often viewed as (and criticized for) being artificial rather than "real." However, a more precise formulation of this criticism is that the experimental situation involves a value of x^N that is not precisely the one in which we are most interested.[18] But given this value, the output is given by $F^M(x^\infty)$ for some x^∞ that matches x^N on the relevant dimensions.

There are many situations x^∞ consistent with an experimental design x^N, as must be the case when we are unable to specify every detail of the experimental situation. One hopes the experimental design determines most of the important aspects of the situation, but cannot control all of the dimensions of the experimental situation x^∞. In effect, the experiment is a *model* of a situation, just as is a theory. The output of an experiment is similarly a model, given by $F^M(x^\infty)$ rather than $F(x^\infty)$. We can hope to identify the salient points of the experimental outcome, but again cannot identify everything.

[16] Hence, the input is drawn from the set $\{x^\infty \in X^\infty : x^\infty(n) = x^N(n),\ n = 1,\ldots,N\}$.

[17] An experiment may yield many observations, but we can arrange the notation to represent the entire experimental outcome as a single observation. This model ignores an issue raised by Roth (1994): the tendency to concentrate on "successes" when reporting experimental results can cause useful information to be neglected. A report of a successful experiment, whether it involves a seeming confirmation or contradiction of a theory, may be less informative than a report that also details the process leading to that experiment. The latter may include investigations of alternative games, alternative experimental procedures, alternative presentations of the experiment to the subjects, and so on. As Roth notes, the line between having also run alternative (possibly unsuccessful) experiments and having run pilot or diagnostic trials is often ambiguous, so that even the best of intentions do not ensure the optimal provision of information. The discussion here assumes that this problem has been solved, so that we have precisely the information we would want from an experiment, and then asks how we combine that information with economic theory.

[18] For example, the experiment may involve university undergraduates choosing between small-stakes lotteries while we may be interested in risk attitudes among large traders in financial markets.

The Goal. The goal of both theoretical and experimental work is to understand the world, or in the context of our model, to understand the function F. How can economic theory, often seemingly quite removed from the world, be combined with experiments in pursuit of this goal?

It is not easy to make this goal more precise. How do we know when we have achieved some understanding of F? We might judge our understanding by the ability to make predictions that match the outputs generated by F. For example, Erev, Roth, Robert L. Slonim, and Greg Barron (2002) pose the following question. Suppose we have both a theory and some experimental evidence bearing on a question of economic behavior, perhaps making different predictions and each potentially subject to error. How do we combine the two to reach a more precise, joint prediction?

The perspective of this paper is different. Instead of asking how we use existing theoretical and experimental results to make predictions, our focus will be on how we can exploit experimental results in the development of more useful theory, and vice versa. We thus shift the emphasis from using existing theory and experiments in making predictions to using them in making new theory and experiments.

To be meaningful, of course, this process must be organized by the ultimate goal of understanding the world. At this point, we confront new difficulties. While we might hope that a theory's predictions will be close, we again cannot expect them to be exact. Then how are we to judge whether a new theory is an improvement? This would be straightforward if there were only benefits and no costs to enhancing the predictive power of a theory, but this is not the case. We return to this issue in section 3.2.

More importantly, the ability to make predictions is only part of what is involved in using economic theory to understand the world. Robert J. Aumann (1985) and Ariel Rubinstein (1998), for example, argue that while the ability to predict behavior may be a good test of our understanding of the world, the ultimate goal is the understanding itself. Economic theory can then be helpful in making precise our intuition and establishing relationships between our ideas, even without adding to our predictive abilities. This is a popular view, but one that makes it all the harder to identify the criteria by which theory is to be evaluated.

3.2 *Why Experiments?*

How do experiments help us assess and design economic theory? It is useful to start by considering the limitations of economic theory, organized around four ideas:

- Economic theory may be inaccurate: given an input $x^N \in X^N$, the theory $f(x^N)$ may produce distributions over outputs that do not match the distribution induced by the environment. Hence, given the information on which it conditions and the results it predicts, the theory provides a result that we would change if we knew the true model.
- Economic theory may be imprecise: the theory may produce a random output of sufficient noisiness to be unhelpful. If possible, we might then seek more precision by increasing N to encompass more information than that captured by X^N, bringing more of the relevant variation in the situation within the purview of the model.
- Economic theory may be uninformative: important information may be missing from the output of the theory. We may then need to expand the range of the theory (increase M).
- Economic theory can be too complicated: if the vectors N and M are large, then the informational demands of the theory may be so burdensome as to make the theory useless.

In practice, we must expect these categories to blur together, with any particular theory exhibiting some degree of each shortcoming.

How can economic experiments help address the shortcomings of economic theory? First, experiments can fill the gap when the theory is either too uninformative or too complicated to be useful.[19] For example, the role of economists in designing and running Federal Communications Commission spectrum auctions in the United States, and subsequently throughout the world, has been offered as evidence for the usefulness of economic theory. Before running the auctions, however, the FCC commissioned experiments (spearheaded by Charles R. Plott) to explore their properties (cf. Paul Milgrom 2004). These experiments played an important role in verifying the internal consistency of the auction procedure and in making the case that the auction could work. Experiments were similarly important in designing the British spectrum auctions (Binmore and Paul Klemperer 2002). There are many other examples, from designing multiunit auctions (Jeffrey S. Banks, John O. Ledyard, and David P. Porter 1989) to designing procedures to allocate access to railroad tracks (Paul J. Brewer and Plott 1996), payload priority on the space shuttle (Ledyard, Porter and Randii Wessen 2002), and airport take-off and landing slots (Stephen J. Rassenti, Vernon L. Smith, and Robert L. Bulfin 1982).

Second, and of more relevance for our discussion, much of the work in experimental economics has centered around identifying inaccuracies and imprecisions in economic theory.[20] For example, the standard economic model of individual behavior is that people maximize expected utility. However, ample experimental evidence suggests that people do not always maximize expected utility, and do not count upon others to do so.[21] More generally, there are long-standing research programs in economics and psychology that serve as a conscience for economic theory, arguing that much of our theory does not provide a good match for behavior.[22]

The difficulty here is that theories are intended to be inaccurate and imprecise. As we have noted, a theory is a deliberate approximation of a world too complicated to be analyzed in complete detail. It is then no surprise to find that the theory does not always match behavior. Experimental confirmation of this fact is potentially helpful, but only if it also points the way toward an improved theory.[23] The constructive role for experiments that challenge economic theories is thus not to simply argue that existing theories do not work, but to point the way to improvements.[24] Perhaps paradoxically, it is when playing this role that experiments pose the greatest challenge to economic theorists. It is relatively easy to dismiss an experimental contention that a theory is sometimes off the mark, but much harder to ignore an indication of how it might be improved.

A new difficulty now appears. When assessing potentially improved theories, we must trade off competing features that leave us with only a partial order over alternatives. Theories are better if they are more accurate and precise, but also if they

[19] This falls into Roth's (1987) category of "Whispering in the Ears of Princes."

[20] This falls into Roth's (1987) category of "Speaking to Theorists."

[21] See Camerer (1995) and Roth (1995) for surveys.

[22] See, for example, Dan Ariely, George Loewenstein, and Drazen Prelec (2003), Daniel Kahneman and Amos Tversky (2000), Loewenstein and Prelec (1992), Richard H. Thaler (1992, 1994), and Tversky and Kahneman (1982).

[23] Neglecting this last point makes it all too easy to fall into a state of tension in which the primary value of experiments is seen as debunking theory, and theory is viewed as having to defend itself from the challenge of experiments. Against this backdrop, it is noteworthy that economic experiments owe much of their prominence to their demonstration that economic theory can be surprisingly robust. For example, Vernon Smith's work on actions (see Theodore Bergstrom [2003] for a survey and Smith [1991, 2000] for collections of papers) showed that elementary supply-and-demand models, the bread-and-butter tool of much of economics, were surprisingly descriptive.

[24] Binmore (1999) advocates such a "consolidating" view of the interaction between theory and experiments.

are more parsimonious (i.e., have smaller N, for fixed M, or in some cases that have smaller N and M). A theory that makes better predictions at the cost of more complexity is not necessarily more desirable. Nor is the goal necessarily a single "correct" theory. Instead, we can expect to work with a portfolio of theories that address different issues and that lie at different points along a frontier that trades off power and complication.

The idea that a more complicated theory may not be better is obscured by economic theory itself, which implicitly assumes that reasoning and inference is costless and automatic.[25] In practice, however, it is a familiar idea that theories are costly to use, and hence that a more accurate or more precise theory is not always superior.[26] This point is often illustrated in introductory economics classes by asking students to think of a road map as a metaphor for an economic theory, and then to note that a map on a scale of 1:1 would be more precise than is commonly found, but its very detail would render it useless.

One of the obstacles to the integration of economic theory and experiments is thus that we have no clear idea of what makes a theory good. For example, we have ample evidence of shortcomings of expected utility theory, as well as an ample collection of alternative models. However, while it is easy to find papers in theory journals working on the tools that might serve as alternatives to expected utility theory, it is much harder to find papers that use these tools. Why? The informal explanation typically is that for most applications, expected utility theory's lack of realism is a reasonable price to pay for its simplicity. This assessment convinces some, while striking others as too easy an excuse.

David W. Harless and Colin F. Camerer (1994) provide a foundation for examining this issue, introducing the notion of an efficiency frontier for generalizations of expected utility theory, balancing predictive power and simplicity. They find that ordinary expected utility theory lies on this frontier, as do several more sophisticated theories. This at least provides some reassurance that expected utility theory is not dominated on every dimension. Depending upon our requirements, we might reasonably choose to work at various points on this frontier, including work with expected utility. But what is the criterion by which points along this frontier are evaluated, other than conventional wisdom and accepted practice? How much of conventional wisdom and accepted practice reflects inertia, historical accident, a lack of familiarity with new theories, fashion, and similar factors? If we are to insist that the goal of a theory is not to be right but to be useful, then one of the great difficulties with economic theory is that we have little consensus on what makes a theory useful, other than that it is customarily used.

This presents a challenge in two respects. Theorists need to be more explicit, both in their theory and in their reactions to experiments, as to how they assess the trade-offs between various limitations. Experimentalists, when interpreting results as supporting an elaboration of existing theory, must address not only the potentially increased precision and accuracy of the theory but also the increased complication.

Is there anything special about experiments in this discussion? In one sense, no. The ideas apply to the use of data in general, regardless of whether an experiment lies

[25] Standard models of reasoning and knowledge begin with a set of states and a partition over these states representing the structure of the available information (e.g., Ronald Fagin, Joseph Y. Halpern, Yoram Moses, and Moshe Y. Vardi 1995). These models have the implication that one automatically knows every implication of any information received. Hence, knowledge of the rules of chess ensures that one knows an optimal strategy, while knowledge of the basic axioms of mathematics makes all of the theorems of mathematics instantly available. It is no surprise that such models do not encourage one to think about the costs of complicated reasoning.

[26] Barton L. Lipman (1999) examines a formal model of the cost of using a theory.

at their source.[27] However, the great attraction and relevance of experiments is the ability they provide to control inputs. If we are interested in assessing the output $f(x^N)$ produced by the theory f in response to input x^N, it may be easier to create (or approximate) input x^N in the laboratory than "in the field."[28] The value of this control becomes all the more apparent upon realizing that we typically can neither ensure that our inputs include all of the factors we would like to have, nor that they exclude all of the ones we would like to not have. At the same time, this advantage brings with it a new challenge. How do we know when the experimental setting has done its job, giving us observations from situations consistent with the desired input x^N and not something else? We return to this issue in section 5.1.

3.3 *Why Theory*

What does economic theory have to contribute? Paralleling the preceding discussion, it is helpful to begin with the limitations of experimental work:

- Experiments may be inaccurate: the experimental procedure is itself a situation. This procedure has presumably been designed to control the key features of the situation, but we cannot expect to have controlled everything. How do we know that the design brings the experimental situation sufficiently close to the real-world situations in which we are interested to be informative about the latter? Experimental psychology refers to this as the question of *external validity*.[29]
- Experiments may be imprecise: our interpretation of an experiment may incorrectly identify the links between the situation and the results. The unrecognized links may make the resulting inferences too noisy to be useful. This is a question of *internal validity*.[30]
- Experiments may be uninformative. It may not be possible to bring the experimental design x^N close enough to the situation to provide useful information.
- Experiments may be informative only at prohibitive cost. Though one of the obvious advantages of experiments is the ability to address otherwise intractable problems in a manageable way, there may be cases where this is not feasible.

Again, these are neither sharply defined nor mutually exclusive categories, and we

[27] The line between experimental and field data is becoming increasingly blurred, as economists turn to "field experiments" designed to capture the best of both settings. See Harrison and John A. List (2004) for an introduction to field experiments and the methodological issues they raise.

[28] If we cannot observe situations consistent with input x^N in the field, why do we care about x^N? The answer is that some values of x^N may provide especially revealing conditions under which to evaluate the theory. For example, theories about bargaining may be more readily evaluated when complicating interpersonal factors are stripped away by examining anonymous bargaining. This in turn my be possible only in the laboratory. For similar reasons, scientists may endeavor to free their experimental environments of impurities, even though such an environment is not observed in nature.

[29] In one view of economic theory, there would be no problem of external validity. Elon Kohlberg and Jean François Mertens (1986, p. 1005) state: "We adhere to the classical point of view that the game under consideration fully describes the real situation—that any (pre)commitment possibilities, any repetitive aspect, any probabilities of error, or any possibility of jointly observing some random event, have already been modeled in the game tree." Pushing this view as far as it will go, the theory then identifies the situation *exactly*. If the theory is simple enough that all of its aspects can be captured in the lab, then we literally have the situation of interest and not simply an approximation, leaving no room for questions of external validity. If not, then the laboratory investigation is irrelevant to the theory. Under this view, an experimental result at odds with the theory tells us only that the experimental design has not captured the conditions under which the theory applies. This classical approach is best viewed as a philosophical exploration of the idea of rationality. It contrasts with a positive approach, under which economic theory is viewed as a tool for modeling and understanding behavior, a tool that is more useful the broader is its applicability. In this case, experimental results at odds with the theory help identify circumstances under which the theory is not applicable.

[30] For example, do the choices of experimental subjects reveal the values they place on the consequences of those choices or some other aspect of the process by which choices are made or values identified? See Harrison, Ronald M. Harstad, and Elisabet Rutström (2002) for a discussion of value elicitation.

can expect experiments to exhibit elements of each.

How can economic theory help? First, economic theory can fill the gap when experiments are not sufficiently informative (at a reasonable cost) to be useful. Oil companies maintain teams of geologists who supplement sampling data with theoretical models designed to predict the likelihood of finding oil beneath a tract of land or ocean bed. Why bother, when a single experimental observation would suffice to provide the result? The difficulty is that the experiment in question consists of drilling a well, which can be sufficiently expensive as to be undertaken only after a favorable theoretical assessment. Similarly, our primary means of assessing nuclear weapons is theoretical, in the form of computer simulations.[31] The relevant experiments are too costly.

Second, and again of more relevance for the current discussion, economic theory can be useful in assessing the external and internal validity of experiments. Insight into links between experimental outcomes and uncontrolled aspects of the experimental situation (and hence external validity), or insight into the link between the experimental environment and the observed behavior (internal validity), can be provided by theoretical models of the behavior. For example, experimental outcomes in continuous double auction markets (e.g., Plott and Smith 1978; Smith 1962, 1964, 1965, 1976, 1982) have been surprisingly efficient, given the apparent thinness of the markets. How do the traders overcome the frictions of a thin market to achieve nearly efficient outcomes? Under what circumstances can we expect similar behavior in actual markets and when should we be less sanguine about efficiency? Addressing the latter question has become easier as theoretical models have tackled the former, showing that the continuous flow of offers, coupled with traders' budget constraints, generates a mechanical but powerful push in the direction of efficient outcomes (Brewer, Maria Huang, Brad Nelson, and Plott 2002; Dhananjay K. Gode and Shyam Sunder 1993, 1993, 1997; Sunder 2004). Alternatively, inconsistent behavior in laboratory decision problems is often interpreted as reflecting preferences that violate the expected-utility axioms. How do we know when we have uncovered something about preferences and when we should seek some other explanation in the experimental design? We have more confidence in the links between behavior and preferences when we have models of the latter.

Once again, a difficulty arises. A model consistent with the observed behavior does not always identify the principles behind the behavior. Instead, experience has shown that economists can build a variety of models consistent with virtually any behavior.[32] How do we know when we have hit upon a clever but irrelevant model and when our model captures something important?[33]

Revisiting a theme, one of the obstacles to the integration of economic theory and

[31] Developing such a theory is itself quite costly, so much so that its provision to other countries is treasonous. But in this case, the relevant experiments involve nuclear detonations whose direct and political costs are even larger.

[32] Difficulties in distinguishing between theories that are consistent with observations and theories that "explain" these observations are not special to economics. Similar considerations arise in the view that one can falsify, but cannot "prove," a scientific theory.

[33] One response to concerns over internal and external validity is to subject the relevant experimental protocol to scrutiny. For example, Thaler (1988) wonders whether Binmore, Shaked, and John Sutton (1985) might have influenced the behavior of their experimental subjects by stressing in their experimental instructions that subjects should maximize their monetary payoffs. As in other experimental sciences, however, a useful response to potential inaccuracy or imprecision in economic experiments is to rely on replication. The more readily an experimental result can be replicated, the less likely is it to hinge upon uncontrolled or unrecognized features of a situation. The evaluation of a new experimental situation then lies in the ability of its "control" treatment to replicate previous results. For two examples among many, Binmore, Shaked, and Sutton begin their experiment by replicating the results of Werner Güth, Rolf Schmittberger, and Bernd Schwarze (1982), and Charles R. Plott and Zeiler (2003) begin their investigation of the endowment effect by replicating previous findings.

experiments is thus that we have no clear idea of when we have a good match between theory and behavior. This difficulty again poses a challenge in two respects. For theorists, there is much to be done in terms of identifying behavior that would enhance one's confidence that the theory in question has captured the relevant principles, or that would force one to question such a conclusion. A good start would be to consistently explain what behavior a theory *cannot* explain.[34] For experimentalists, it can be important to argue not only that a model captures the outcomes of the experiment, but that it captures the appropriate links between the experimental situation and the outcome. Again, a good start would be to consistently explain what outcomes would lead to the opposite conclusion.

4. *Combining Theory and Experiments*

4.1 *Using Experiments to Learn About Theory*

4.1.1 *Testing Theory: Accuracy*

How can we use experiments to evaluate economic theory? Suppose we fix an experimental design x^N and a set of possible outputs S^M, identifying the features of the input and output that are considered salient in the experiment. The resulting experiment produces an output s^M. Does this indicate that we should be more confident of economic theories that place relatively large probability on the outcome s^M, or on similar outcomes, when faced with the input x^N? Some useful insight into this question is given by the following argument, adapted from Alvaro Sandroni (2002), that is typical of the calibration literature.

Given the design x^N, the experiment's output s^M is randomly determined by the environment. In particular, a situation x^∞ is randomly drawn from the set of situations whose first N dimensions match x^N, i.e., from the set of situations that match the experimental design in those features controlled by the design. This situation is then converted into an output according to the function F describing the environment, and we observe the first M dimensions of this output, giving the output s^M. We let π^* denote the resulting probability distribution over the set of possible experimental outputs S^M, and refer to π^* as the *true* distribution.[35]

Similarly, given the input x^N, a theory f can be viewed as producing an output $\pi \in \Delta S^M$, i.e., a probability distribution over the set of possible outcomes . This output is itself randomly chosen according to a probability distribution over ΔS^M that is determined by the theory.[36] We let f^* denote this distribution.

The task now is to describe the implications of the experiment for the theory. We think of running the experiment, producing a randomly-drawn output s^M (from the distribution π^* induced by the experiment), and choosing a randomly-drawn distribution π (from the distribution f^* induced by the theory). We insert these realizations into an *evaluation rule* $T(s^M,\pi)$. The evaluation rule produces the output $T(s^M,\pi)=1$ if we accept the theory given realizations π and s^M and $T(s^M,\pi)=0$ if we reject the theory given realizations π and s^M. Clearly, of course, a single experiment does not suffice to evaluate a theory. The labels "accept" and "reject" might accordingly be more precisely (but also more cumbersomely) phrased as "regard this experiment as evidence in favor

[34] Among many such examples, Timothy N. Cason and Daniel Friedman (1996) and John H. Kagel, Harstad, and Dan Levin (1987) begin their analysis with theoretical models, focusing on aspects of behavior the models cannot accommodate.

[35] Formally, $\pi^*(s^M)$ is proportional (being rescaled to ensure a total probability of one) to $\rho(\{x^\infty \in X^\infty : x^\infty(n) = x^N(n),\ n = 1,\ldots, N \text{ and } F^M(x^\infty) = s^M\})$.

[36] Recall that a theory is an element of $\Delta\Delta S^M$, being a distribution from which a distribution over s^M is randomly drawn. Notice that the outcomes of the experiment and the theory are drawn from different spaces. This is familiar. For example, the experiment produces the outcome *rain* or *no rain*, according to a distribution that depends upon such factors as the location and the season. The theory is allowed to announce a probability of rain, which may itself be drawn from a distribution that depends upon similar factors.

of the theory" and "regard this experiment as evidence questioning the theory."

How do we design a useful evaluation rule T? One desirable criterion is that if one were to offer the true distribution π^* as the realized output of one's theory, then our evaluation of the experimental evidence should be unlikely to reject it. Because the experimental outcome is random, we cannot expect the distribution π^* to always prompt an acceptance. For example, the evidence will sometimes reject the theory that a fair coin yields heads on half of its flips, simply because we encounter an unusual and unlikely sequence of outcomes. However, we can reasonably ask that such rejections be rare. We make this idea precise by saying that an evaluation rule *accepts the truth with probability at least* $1-\epsilon$ if, for *any* true distribution π^*,

$$\pi^*\left(\left\{s^M : T\left(s^M, \pi^*\right)=1\right\}\right) \geq 1-\epsilon.$$

Hence, with probability at least $1-\epsilon$, the true distribution π^* generates an experimental outcome s^M that would not prompt us to reject the truth, if we were asked to evaluate the truth as a possible theory. Notice that we will typically not know the true distribution π^* when designing an evaluation rule, and hence our requirement is that the evaluation rule be unlikely to reject the truth (given the distribution of experimental outcomes generated by the truth), regardless of what the truth happens to be.[37]

At the other end of the spectrum, an evaluation rule is not particularly helpful in assessing a theory if there are *no* experimental outcomes that would cause the theory to be rejected (even though this would be one way to accept the truth with high probability). For example, an experimental test of the theory that a coin is fair is not helpful if it always accepts the theory, but could be useful if it instead rejects the theory if the observed proportion of heads (or tails) is too large. To capture this distinction, we say that an evaluation rule is *blindly passed* by theory f with probability $1-\epsilon$ if, for *every* $s^M \in S^M$,

$$f^*\left(\left\{\pi : T\left(s^M, \pi\right)=1\right\}\right) \geq 1-\epsilon.$$

Hence, no matter what observation s^M the experiment produces, with probability at least $1-\epsilon$ the theory f (via its induced distribution f^*) produces a distribution π over possible experimental outcomes that causes the theory to be accepted (given the observation s^M). The phrase "blindly passed" here refers to the fact that the theory f is accepted by the evaluation rule with probability at least $1-\epsilon$ *regardless of the experimental outcome* or, equivalently, to the fact that f embodies no understanding of the true process generating experimental outcomes. As a result, a theory may blindly pass an evaluation rule with high probability, but without providing any insight into the principles governing the outcome in this situation.

The main result (proven in section 7) is now:[38]

Proposition 1 *Any evaluation rule that accepts the truth with probability* $1-\epsilon$ *can be blindly passed with probability* $1-\epsilon$.

At first glance, it seems obvious that an evaluation rule that accepts the truth can be passed—one need only propose the truth as one's theory. However, Proposition 1 makes a quite different assertion. If an evaluation rule accepts the truth sufficiently often (i.e., with probability $1-\epsilon$), then one can find a theory that requires no knowledge of the truth and has the property that, no matter what the outcome of the experiment and no matter what the actual process generating the experimental outcomes, the theory is accepted with probability $1-\epsilon$. The following illustrates:

[37] For example, we can design an evaluation rule that can observe one hundred flips of a coin and simultaneously be quite likely to conclude that the coin is biased towards heads when it is, and quite likely to conclude that the coin is biased toward tails when it is, because these two biases (if true) generate quite different distributions over experimental outcomes.

[38] This is a special case of Proposition 1 in Alvaro Sandroni (2002).

Example. Suppose that there are only two possible outcomes of an experiment, *head* and *tail*. The environment induces a true probability distribution over these two outcomes, which we denote as $\pi^* \in [0,1]$, where π^* is the probability of the experimental outcome *head*. As the notation suggests, we can think of the experiment as a single flip of a (possibly biased) coin, with π^* being the true probability of a *head*. The theory generates a (possibly randomly determined) candidate probability π, which we must then combine with the experimental outcome to evaluate the theory. A possible evaluation rule is:

$$T(s^M,\pi)=\begin{cases}1 & \text{if } s^M = \textit{tail} \text{ and } \pi < \frac{1}{3} \\ 1 & \text{if } s^M = \textit{head} \text{ and } \pi \geq \frac{1}{3} \\ 0 & \text{otherwise}\end{cases}.$$

Hence, the theory is accepted if the experimental realization is *tail* and the realization π of the theory attaches probability less than $\frac{1}{3}$ to *head* (the first line), and is accepted if the experimental realization is *head* and the realization of the theory attaches probability at least $\frac{1}{3}$ to *head* (the second line). This particular evaluation rule accepts the truth with probability at least $\frac{1}{3}$.[39] Such a minimum acceptance probability does not sound very impressive. By altering the evaluation rule, we could manage to boost this probability to $\frac{1}{2}$, but could not go further in this case.[40] Now suppose the theory f draws π uniformly from the set [0,1]. Hence, consistent with the model of Section 3.1, the probability π with which the theory predicts the outcome *head* is itself drawn randomly according to a distribution f^* over ΔS^M. Given the uniform distribution we clearly work without any information as to what the truth might be. Then

$$f^*(\{\pi : T(tail,\pi)=1\}) = f^*(\{\pi < \tfrac{1}{3}\}) = \tfrac{1}{3}.$$

$$f^*(\{\pi : T(head,\pi)=1\}) = f^*(\{\pi \geq \tfrac{1}{3}\}) = \tfrac{2}{3},$$

and hence the evaluation rule is blindly passed with probability $\frac{1}{3}$.

To see the intuition behind Proposition 1, think of playing a zero-sum game against a malevolent and possibly omniscient opponent, "Nature," where Nature chooses the true theory π^* generating the experimental outcomes and you choose a theory f, with Nature attempting to maximize the probability of an outcome that rejects your theory (here we see Nature's malevolence) and you trying to minimize this probability. Suppose (counterfactually) that you had the luxury of observing Nature's choice before making your own. Then you could always simply name Nature's choice as your theory, and the requirement that the test accept the truth with probability $1-\epsilon$ ensures that your success probability would be at least $1-\epsilon$. Alternatively, the worst that could happen is that Nature gets to observe your proposed theory before choosing the truth (here we see Nature's potential omniscience) and then chooses the truth to minimize your success rate.[41] The minmax theorem then gives us a result that is familiar in the context of zero-sum games, namely that you can do as well in the second circumstance as in the first, and hence can succeed with probability at least $1-\epsilon$ in the second circumstance. But your optimal performance in the actual game, in which neither side gets to observe the other's move, must be somewhere between these best and worst cases, ensuring that the test can be blindly passed with probability $1-\epsilon$.

[39] If the true distribution is $\pi^* < \frac{1}{3}$, the evaluation rule accepts π^* if the experiment generates outcome *tail*, which happens with probability $1-\pi^* (> \frac{1}{3})$. If the true distribution is $\pi^* > \frac{1}{3}$, the evaluation rule accepts π^* if the experiment generates outcome *head*, which happens with probability $\pi^* (> \frac{1}{3})$.

[40] It is to be expected that an experiment with only two outcomes provides rather crude information—how much information can one expect to extract about the probability of heads, from a coin of unknown bias, from a single flip? Higher minimum acceptance probabilities require richer outcome spaces. Whether we are better off in this case with a rule that accepts the truth with probability $\frac{1}{2}$ depends upon the relative costs of mistakenly accepting or rejecting the various values of π.

[41] Here, it is clear that one is not simply predicting well by offering the truth as a prediction, since the prediction is chosen first and then a worst-case specification of the truth is chosen.

The implication of this result is that the ability of an economic theory to match experimental data does not necessarily provide evidence in support of the theory. Instead, given any specification of questions that a theory could be asked, and any specification of how the answers to these questions are to be compared to the experimental evidence, one can devise a theory based on no understanding of the situation or the underlying principles that allows one to be as successful as knowing those principles precisely.

This result is not simply a restatement of the common view that it is somehow more instructive if one first commits to a theory and then compares it to data (rather than first observing the data and then constructing a theoretical rationalization). More importantly, this result is not simply a restatement of the observation that it is important for theories constructed in response to experimental observations to make "out of sample" predictions, i.e., predictions that could be assessed only with the collection of new data.

Instead, the ability to blindly construct a theory f that fares as well as the truth depends upon knowing the *evaluation rule T* by which the theory is to be assessed. As long as we identify a fixed set of potential tests to which a theory is to be subjected, whether in or out of sample, we can blindly construct a theory that fares as well as the truth in these tests, regardless of whether we have seen the outcomes of the tests and regardless of what these outcomes might be. Interpreting experimental evidence as supporting a theory, or offering a theory as an interpretation of experimental evidence, thus acquires some bite only if the theory is clear and complete enough that it can be extended to answer new questions and confront new tests that did *not* play a role in the construction of the theory.[42] Is the theory clear enough that others could design new tests, and is one willing to risk the theory in such tests? If not, then it is not clear that progress has been made.

For example, Bolton and Ockenfels (2000) and Fehr and Schmidt (1999) offer models motivated by behavior in bargaining experiments, with each model consisting of an explicit specification of how utility depends upon (one's own and one's rival's) payoffs (cf. section 2). In doing so, the authors are offering models that (like all others) cannot hope to capture every detail of human motivation, and hence are bound to fail some tests. However, these models exhibit the essential characteristic of being sufficiently precise and powerful that new tests can be devised. The authors are taking some risk in presenting their theories so explicitly, but in return they ensure that their models can be meaningfully investigated experimentally. If their models do not provide useful alternatives to the hypothesis that players maximize their expected monetary payoffs, they will be stepping stones to such alternatives. Either way, their models allow progress that would be impossible without the ability to venture beyond the experimental designs that prompted them.

4.1.2 *The Margin of Error: Precision*

We have modeled a theory as producing a probability distribution over probability distributions over outcomes.[43] In most cases, an economic theory provides nothing of the sort, with deterministic outcomes being the rule. How do we put these two together?

Think first about how economists typically do empirical work. The underlying intuition and theoretical structure come from a model free of anything random. But before confronting this model with the

[42] Eddie Dekel and Yossi Feinberg (2004) propose a test for whether one's theory matches the environmental function F that hinges upon asking one to design (rather than react to) an evaluation rule T.

[43] This ability to mix is important, as without it one cannot be assured of blindly passing evaluation rules that accept the truth.

data generated by a noisy world, an error term is added. The characteristics of this error can be important, providing the foundations for the inferences to be drawn from the results.

Assumptions about errors play a similarly important role in interpreting experimental results. One argues not that the data and the theory are a perfect match, but rather that the errors required to reconcile the data with the model are not too large.

What does "not too large" mean? Auctions have received significant attention from experimentalists, with results that often appear to be at odds with theoretical predictions.[44] One interpretation of the observed behavior is to assume that subjects invariably intend to identify and take their optimal actions, but that some sort of "tremble" translates this optimal action into a random choice.[45] The evidence convinces most observers that by this standard, there is often a large gap between theoretical results and experimental behavior: the trembles required to reconcile the two are too large, and hence much of auction theory appears insufficiently accurate to be a useful description of behavior.

Alternatively, one might interpret the observed behavior by assuming that subjects are only ϵ-optimizers, being content with identifying and playing an action that is within some ϵ of a best response. Section 2 touched on Harrison's (1989, 1992) argument that, by this standard, very little error is required to reconcile the theory with the data. It turns out that one's actions have relatively little effect on expected payoffs in many auctions (as long as actions are not too far from equilibrium), and hence that one can come close to maximizing one's expected payoff with actions that seem far away from the equilibrium. If we are to view errors in this way, then we must be careful in concluding either that optimization is a poor description of individual behavior or that the outcome is not (approximately) in equilibrium.[46]

Yet another interpretation of the observed behavior assumes that subjects choose their actions not by optimizing but through a process of trial-and-error learning.[47] Here, errors are measured in terms of the strength of the incentives embedded in the learning process.

The implication in each case is that the interpretation of experimental results requires not only a theory, but also some idea of what types of errors are most likely involved when the theory does not work perfectly. Richard D. McKelvey and Thomas R. Palfrey's (1995) *quantal response equilibrium* is perhaps the best developed and most general such model, built around agents who maximize utility functions perturbed by random terms. Notice that the errors here are built into the model of individual behavior from the beginning rather than being added at the end.[48] These errors can be interpreted as capturing unmodeled but (one hopes) small effects on preferences. Quantal response equilibria have been used to good

[44] For surveys, see Douglas D. Davis and Charles A. Holt (1993), Kagel (1995), and Kagel and Levin (2002).

[45] Such trembles may initially appear difficult to motivate, but similar ideas have played an important role in the equilibrium refinements literature. More importantly, the possibility that typing or other errors might lead to mistaken bids was a serious concern in the design of the FCC spectrum auctions (Paul Milgrom 2004).

[46] At the same time, the experimental results still present a challenge for the theory. We no longer have evidence that the model is inaccurate, but we have evidence that it is not sufficiently precise to be a useful description of *behavior*. In response, we could restrict our attention to payoffs (effectively, shortening the list M of outputs of the theory) or refine the theory in hopes of more precisely capturing behavior.

[47] Binmore and Samuelson (1999) study learning models whose results depend importantly on the nature of (possibly very small) errors.

[48] It is a familiar result that incorporating uncertainty into the construction of a model can yield results that differ from simply appending error terms to a deterministic model. For example, incorporating an error term into players' choices in a game and then solving for a (perfect) equilibrium (Reinhard Selten 1975) can give results quite different than first solving for a (Nash) equilibrium and then adding an error term.

effect in analyzing a variety of experimental results.[49]

Once again, however, new challenges appear. Philip Haile, Ali Hortacsu, and Grigory Kosenok (2004) show that quantal response equilibrium is a sufficiently flexible notion that, by appropriately specifying the error terms, one can obtain equilibria consistent with *any* behavior that one might possibly observe. The unmodeled errors are thus important. Without further assumptions concerning their distribution, too much is left out of the model for its predictions to be usefully precise.[50]

There are then two possibilities for harnessing the potential power of quantal response equilibria. First, quantal response models can provide comparative static implications even without distributional assumptions.[51] Alternatively, Haile, Hortacsu, and Kosenok's (2004) result depends upon having sufficient freedom in specifying the errors in the individual utilities underlying the quantal response model. We may often have either intuition or experimental evidence about what forces are captured by the errors. We may then augment the underlying model with hypotheses about the distribution of errors sufficiently powerful to produce precise results. In effect, we are enhancing precision by expanding the set of inputs X^N to capture more information. Jacob K. Goeree, Holt, and Palfrey (2004) note that applications of quantal response equilibria typically work with models that are monotonic, in the sense that increasing the expected payoff of an alternative increases the probability that it is chosen.[52] Goeree, Holt, and Palfrey provide sufficient conditions for quantal response equilibria to be monotonic and show that monotonic quantal response equilibria can have substantive empirical content.[53] The informativeness of experiments based on quantal response models is thus enhanced by a better theoretical understanding of such models.

Focusing attention on the specification of errors has the advantage of leading naturally to a provision for heterogeneity in players' behavior. Perhaps one of the most robust findings to emerge from experimental economics is that such heterogeneity is widespread and substantial. Despite this, heterogeneity has often not played a prominent role in many theoretical models. Instead, theoretical explanations often have the flavor of seeking "the" model of individual behavior that will account for the experimental behavior. This appears to be a holdover from the original presumption that monetary payoffs, common to all subjects, suitably captured preferences, an assumption that encourages a view of players as homogeneous.[54] Error terms provide a natural vehicle for capturing heterogeneity.

The implication is that there is much to be gained by making our treatment of errors in individual decision-making more explicit, and hence much to be gained in the interpretation of experimental results by being more careful with our theory. However, this is a task made all the more daunting by the

[49] See, for example, Simon P. Anderson, Jacob K. Goeree, and Charles A. Holt (1998, 1998, 2001), Georee and Holt (2001), Goeree, Holt and Thomas R. Palfrey (2002), and Richard D. McKelvey and Palfrey (1992, 1995, 1998).

[50] Though the technical details are different, this result is similar in spirit to John O. Ledyard's (1986) observation that any behavior is consistent with the notion of Bayesian equilibrium.

[51] A concentration on comparative statics requires that the distributions of the error distributions do not vary (or vary sufficiently regularly) as the parameters of the problem vary, a requirement lying behind many an econometric inquiry.

[52] For example, a logit choice model with independent, identically distributed extreme-value errors satisfies monotonicity.

[53] Other examples of work focussing on the structure of errors include Mahmoud A. El-Gamal and David M. Grether (1995), David W. Harless and Camerer (1995), Harrison (1990), and Daniel Houser, Michael Keane, and McCabe (1995). Similarly, Ledyard's (1986) analysis of Bayesian equilibrium suggests that we augment the model, perhaps with assumptions about players' beliefs.

[54] For work on subject heterogeneity, see Andreoni, Marco Castillo, and Ragan Petrie (2003), Andreoni and John Miller (2002), and the examples cited in note 53.

observation that the considerations relegated to error terms are often there because we know little about them. Once again, theorists are sent back to the drawing board in search of theories precise enough to be useful.

4.2 *Using Theory to Learn about Experiments*

4.2.1 *External Validity*

Having found an experimental regularity, how do we assess whether the experimental design from which it emerges is a good match for the intended application (the question of eternal validity) and whether we have linked the resulting behavior to the appropriate characteristics of the design? The obvious observation is that more experiments are always helpful, and one of the great advantages of the experimental method is the ability to collect more data. But economic theory has a role to play in conjunction with these experiments.

For example, the standard assumption when modeling intertemporal choice is that people maximize the sum of exponentially-discounted expected utilities. Expected utility theory derives much of its appeal from the fact that it rests upon a collection of axioms that can be interpreted as prescribing consistent behavior (Leonard J. Savage 1972). Extending this argument to intertemporal behavior, consistency is similarly ensured by exponential discounting.

The difficulty is that the experimental evidence has not been particularly supportive of exponential discounting.[55] The consensus leans toward a model in which discounting departs from exponential in the direction of being biased toward the present, so that discount rates decline as one evaluates more distant payoffs. Hyperbolic discounting is the most prominent example.

The case for hyperbolic discounting (or other forms involving a bias toward the present) is often bolstered with results from (nonhuman) animal as well as human experiments. The use of hyperbolic discounting in interpreting results from animal experiments is routine (e.g., James E. Mazur 1984, 1986, 1987). The question to be considered here is one of external validity: how relevant is the animal evidence for human behavior?

Our approach to this question is not to debate how similar are animals and humans, but rather how similar are the typical discounting problems faced by animals and humans. In turn, the approach to this latter question is to examine theoretical models of these discounting problems.

Discounting in animals is commonly examined in the context of foraging behavior (e.g., Alasdair I. Houston and John M. McNamara (1999), Alex Kacelnik (1997), Michael Bulmer (1997)). It is helpful to begin with a highly simplified, deterministic model. Suppose an animal faces the problem of maximizing total food consumption over an interval of length T. A function $c:\mathbb{R}_+\rightarrow\mathbb{R}_+$ identifies the quantity of consumption $c(t)$ that can be secured upon the investment of foraging time t. The animal is to make a succession of foraging-time/consumption pairs of the form $(t,c(t))$, where each choice $(t,c(t))$ allows consumption $c(t)$ but precludes another choice until time t has passed.

The animal's task is to choose an optimal pair $(t_r,c(t_r))$ for any length τ of time remaining in the foraging interval. Let $V(\tau)$ be the value of the optimal continuation consumption plan, given the length τ of time remaining.[56] If τ is sufficiently large, then the optimal consumption plan will be nearly stationary, featuring a choice of some fixed, optimal t^* at each opportunity. This allows the approximation

$$V(\tau)=\tau\frac{c(t^*)}{t^*}.$$

[55] See Shane Frederick, Loewenstein, and Ted O'Donoghue (2002) for a survey and Maribeth Coller, Harrison, and Rutström (2003) for an alternative view.

[56] The function V is implicitly defined by $t_\tau\in\arg\max_t\{c(t)+V(\tau-t)\}$.

But then the optimal consumption plan t^* maximizes

$$\frac{c(t)}{t} \tag{1}$$

Hence, optimal foraging behavior induces a preference for consumption $c(t)$ at time t over $c(t')$ at t' if

$$\frac{c(t)}{t} > \frac{c(t')}{t'}.$$

Consumption at time t is thus optimally discounted by $1/t$, i.e., is discounted by the hyperbolic function $1/t$. It then seems unsurprising that experiments with animals are suggestive of hyperbolic discounting.

How relevant is this evidence for humans? Hyperbolic discounting arises out of a model in which delayed consumption imposes an opportunity cost, in the sense that other consumption opportunities are precluded while waiting for the current realization. The variable t measures the time spent foraging, during which consumption is precluded. There is nothing like this in the intertemporal decision problems typically associated with hyperbolic discounting in humans, where t measures a delay during which other options are not closed. For example, when facing the canonical hyperbolic-discounting story of choosing between one sum of money now and another in a week, and then between the same sums in fifty-two and fifty-three weeks, there is no presumption that intervening consumption possibilities are sacrificed. We thus have reason to doubt that hyperbolic discounting in animals has sufficient external validity to be of relevance for human behavior.

This observation is only the first step of the story. There may still be good reasons for humans to engage in hyperbolic discounting. One possibility is that human intertemporal preferences were formed during a time in which people typically faced decision problems similar to the foraging choices thought to be typical of animal decisions, and that people now simply apply the resulting (hyperbolically discounted) preferences to current decisions without noting the different context.[57] In effect, the opportunity costs of the time sacrificed while waiting for consumption may have been important in the ordinary lives of our ancestors, even if we do not commonly encounter it in our lives, potentially restoring the relevance of the animal experiments.[58]

A second difficulty now arises. Suppose we expand our simple foraging model to accommodate uncertainty. Let $\{X(1),\ldots, X(n)\}$ be a collection of independent, positive-valued random variables. We interpret each of these as representing a foraging strategy, with each foraging strategy characterized by a random length of time until it yields a consumption opportunity. To keep the example transparent, we simplify our previous model by assuming that each consumption opportunity features one unit of food. The animal chooses a foraging strategy, waits until its payoff is realized, chooses another strategy (perhaps the same one), and so on, until a fixed foraging period of length T has been exhausted.

This model gives what is commonly known as a renewal process. The intuition is that once a unit of food has been received, the process has been "renewed," in the sense that the set of possible choices and outcomes has reverted (literally in the case of an infinite horizon and approximately in the case of a sufficiently long finite horizon) to its original configuration. For sufficiently long horizons, the optimal strategy will again be approximately stationary. Consider a stationary strategy, in which the same random variable $X(i)$ is chosen at each opportunity. Let μ_i be the mean time before food is realized under $X(i)$.

[57] Peter D. Sozou (1998) and Partha Dasgupta and Eric Maskin (2003) explore evolutionary motivations for hyperbolic discounting that do *not* depend upon foraging as the standard decision problem.

[58] This possibility provides one illustration of how elusive external validity can be. There is often no single or obvious external situation to which the model is to be applied. The question may then not be whether there are situations outside the laboratory that correspond to the experiment, but rather whether the corresponding situations are the "right" ones. We return to this point at the end of this section.

Let $N(T)$ be the number of renewals (i.e., number of units of food) secured by time T. Then the elementary renewal theorem (Sheldon Ross (1996, Proposition 3.3.1)) indicates that, as T gets large,

$$\frac{N(T)}{T} \to \frac{1}{\mu_i} \quad .$$

As a result, the stationary strategy that chooses the random variable with the smallest mean time to renewal (μ_i) will be approximately optimal (among the set of all strategies, not just stationary ones), in the sense that it maximizes the number of renewals $N(T)$ and hence consumption, for large values of T. This strategy chooses the random variable $X(i)$ that maximizes

$$\frac{1}{E\{t\}}, \tag{2}$$

where t is the renewal time and $E\{t\}=\mu_i$ is its expected value. In contrast, applications of hyperbolic discounting in economics typically assume that people maximize the expected value of hyperbolically-discounted utilities. In our simplified case, recalling that each random delay is terminated by the appearance of one unit of food, this calls for maximizing

$$E\{1/t\}, \tag{3}$$

The objectives given by (2) and (3) can especially differ if the menu of foraging strategies includes alternatives with high mean renewal times but that attach some probability to very short waiting times. Such strategies may fare very well under (3), while being less attractive under (2).

We thus find that an appeal to our evolutionary background may or may not allow us to interpret animal evidence as bracing a belief in human hyperbolic discounting, but that in the process we also provide evidence against commonly-used models of (hyperbolically discounted) expected utility maximization. There appears to be no obvious way to interpret animal experiments as supporting both hyperbolic discounting and expected utility maximization.

Two qualifications are relevant. First, there are things about animal behavior that we do not understand.[59] More importantly, the point here is not to defend exponential discounting. Instead, it would be quite a surprise if discounting were precisely exponential. There is also evidence of hyperbolic discounting from human experiments, which the current discussion does not call into question.[60] The point is that extending results from animal experiments to conclusions about human behavior raises questions of external validity that can be examined through the lens of economic theory. In connection with hyperbolic discounting, the accompanying theory is not immediately supportive of a link.

Assessments of external validity can be further complicated by the fact that the appropriate external environment for comparison is often not obvious. Consider one of the simplest experimental settings, the dictator game. Experiments find that dictators typically do not seize all of the money, despite the lack of any obvious reason for not doing so (Davis and Holt 1993; Robert Forsythe, Joel L. Horowitz, N. E. Savin, and Martin Sefton 1995). What should we make of this result? Each of us is constantly involved in a version of the dictator game, in that we constantly have opportunities to give away the money in our wallets, or anything else that we own. Typically, however, we hold on to what is ours. One might then view the experimental evidence as being swamped by a mass of practical experience with the dictator game, in which people for the most part tend to keep what they have.

[59] One of the puzzles facing biologists is that observed behavior appears to match the objective given by (3) more closely than the simple theoretical prediction that (2) be maximized (Melissa Bateson and Alex Kacelnik (1996), Kacelnik (1997), Kacelnik and Fausto Brito e Abreu (1998)).

[60] Again, see Frederick, Loewenstein, and O'Donoghue (2002). Here, as always, there are still questions of internal validity. Is the observed behavior a product of hyperbolic discounting, or something else? Halevy (2004) and Ariel Rubinstein (2003) explore alternatives.

Then what do the experiments have to tell us? One message is clear: people do not always keep everything. This is a useful point of departure. Outside the laboratory, people also sometimes relinquish what they own, giving gifts and making contributions to charity. A variety of explanations have been offered for why this seemingly altruistic behavior is consistent with rational, selfish behavior.[61] While often persuasive, and consistent with some aspects of behavior in dictator experiments,[62] it seems a stretch to suggest that such explanations can cover every bit of generosity. One can then view dictator experiments as an attempt to strip away the confounding factors and isolate a situation in which rational, selfish behavior has a clear prediction, allowing us to conclude that people are not *always* relentlessly selfish.

This is instructive, but only the most extreme would claim that selfish preferences are a complete description in every circumstance. Do the dictator experiments have anything to contribute beyond challenging such extremists? Here we return squarely to the question of what is the appropriate context in which to evaluate the external validity of dictator experiments. Does the experimental allocation represent the continual decisions we implicitly make about whether to keep our wealth or give it away? If so, then the findings provide a serious challenge to the preferences commonly used in economic models. Does the experiment capture those rarer circumstances under which people make anonymous contributions to charity? If so, then the findings are commonplace. A useful point of departure in addressing this issue is again theoretical, aimed at identifying and modeling the features that distinguish the first set of circumstances from the second, and then interpreting these circumstances in terms of experimental designs and findings. Once again, the general point is that examining the relevant theory can help assess the interpretation and external validity of experimental results.

[61] For example, people are said to give gifts in anticipation of reciprocation, to contribute to charity in order to gain esteem, to tip in order to advertise their generosity to fellow diners, and so on.

[62] For example, the sensitivity of amounts retained by dictators to the degree of anonymity in the experiment (Hoffman, McCabe, Keith Shachat, and Smith 1994; Hoffman, McCabe, and Smith 1996) could be interpreted as indicating that one purpose of a seemingly altruistic act is to demonstrate one's behavior to others.

4.2.2 *Internal Validity*

Experiments in economics typically feature monetary payoffs. Can we assume that these monetary payoffs represent utilities? Section 2 touched on one reason why the answer might be no, namely that subjects might care about more than simply the amount of money they make. However, suppose that this is not the case. If subjects are risk averse, then monetary payoffs still do not provide a good representation of utility.

One of the early insights of experimental economics was that we can effectively eliminate risk aversion, as long as subjects are expected-utility maximizers. Suppose one has in mind an experiment that would make monetary payments ranging from 0 to 100. Then replace each payoff $x \in [0,100]$ with a lottery that offers a prize of 100 with probability $x/100$ and a prize of zero otherwise. Expected payoffs are unchanged. However, for any expected utility maximizer, regardless of risk attitudes, the expected utility of a lottery that pays 100 with probability p (and 0 otherwise) is

$$pU(100)+(1-p)U(0)=U(0)+[U(100)-U(0)]p.$$

This expression is linear in p, meaning that the agent is risk neutral in the currency of probabilities. On the strength of this convenience, lottery payoffs have often been used in experimental economics.[63]

[63] See Cedric A. B. Smith (1961) for an early theoretical discussion of lottery payoffs, and Roth and Michael W. K. Malouf (1979), Roth and J. Keith Murnighan (1982), and Roth and Françoise Schoumaker (1983) for early experimental applications.

Against this background, Matthew Rabin (2000) (see also Rabin and Richard H. Thaler 2001) presents an argument that we illustrate with the following example. Suppose Alice would rather take \$95.00 with certainty than face a lottery that pays nothing with probability $\frac{1}{2}$ and \$200 with probability $\frac{1}{2}$. Suppose further that Alice would make this choice no matter what her wealth. Then either the standard model of utility maximization does not apply, or Alice is absurdly risk averse.

To see the reasoning behind this argument, assume that Alice has a differentiable utility function $U(w)$ over her level of wealth w, with (at least weakly) decreasing marginal utility. Alice's choice implies that the utility of an extra 95 dollars is more than half the utility of an extra 200 dollars. This implies that $\frac{1}{2}200U'(w+200)<95U'(w)$, where w is Alice's current wealth and $U'(w)$ is the largest marginal utility found in the interval $[w, w+200]$ and $U'(w+200)$ is the smallest marginal utility in that interval.[64] Simplifying, we have, for any wealth w

$$U'(w+200)\leq\frac{19}{20}U'(w). \tag{4}$$

Now letting w_0 be Alice's initial wealth level and stringing such inequalities together, it follows that, for any w, no matter how large, Alice's utility $U(w)$ satisfies

$$\begin{aligned}U(w) &\leq U(w_0)+200U'(w_0)+200U'(w_0+200)+200U'(w_0+400)\\ &\quad +200U'(w_0+600)+\ldots\\ &\leq U(w_0)+200\left[U'(w_0)+\frac{19}{20}U'(w_0)+\left(\frac{19}{20}\right)^2U'(w_0)+\ldots\right]\\ &= U(w_0)+200U'(w_0)/\left(1-\frac{19}{20}\right)\\ &= U(w_0)+4000U'(w_0).\end{aligned}$$

[64] Hence, $95U'(w)$ is an upper bound on the utility of an extra 95 dollars, and $200U'(w+200)$ a lower bound on the utility of an extra 200 dollars. Rabin (2000) contains additional examples and shows that the argument extends beyond the particular formulation presented here.

where the first inequality breaks $[0,\infty]$ into intervals of length 200 and assumes that the maximum possible marginal utility holds throughout each interval, the second repeatedly uses (4), and the remainder is a straightforward calculation.

Now consider a loss of 3000. A similar argument shows that the utility $U(w_0-3000)$ must satisfy

$$\begin{aligned}U(w_0-3000) &\leq U(w_0)-200U'(w_0)-200U'(w_0-200)-\ldots-200U'(w_0-2800)\\ &\leq U(w_0)-200\left[U'(w_0)+\frac{20}{19}\times U'(w_0)+\ldots+\left(\frac{20}{19}\right)^{14}U'(w_0)\right]\\ &\leq U(w_0)-4400U'(w_0).\end{aligned}$$

Comparing these two results, we have that for any $X>0$,

$$\frac{1}{2}U(w_0-3000)+\frac{1}{2}U(w_0+X)\leq U(w_0)-400U'(w_0)<U(w_0).$$

Hence, there is no positive amount of money X, no matter how large, that would induce Alice, no matter how wealthy, to accept a fifty/fifty lottery of losing 3000 and winning X. Risk aversion over relatively small stakes thus implies absurd risk aversion over larger stakes.

Risk aversion over small stakes seems quite reasonable and is consistent with laboratory evidence (e.g., Holt and Susan K. Laury 2002). How do we reconcile this with the seeming absurdity of the implied behavior over larger stakes? Taking it for granted that people are not so risk averse over large stakes, Rabin (2000) and Rabin and Thaler (2001) suggest that the expected-utility model should be abandoned.

This conclusion poses a puzzle for experimental practice. The use of lottery payoffs appears to be either unnecessary (because

subjects are risk neutral over the relatively modest sums paid in experiments) or necessarily ineffective (because subjects are risk averse over small sums, and hence cannot be expected-utility maximizers). The argument is even more challenging for economic theory, where expected-utility maximization is firmly entrenched.

In response, our attention turns to questions of internal validity. Is the observed behavior appropriately interpreted as reflecting departures from expected-utility maximization? Addressing this question requires a more careful look at the theory. Let X be a set of consequences, Ω a set of states, and $\mathcal{L}$ a set of *acts*, where an act is a function associating a consequence with each state. Savage (1972) shows that if an agent has preferences over the set of acts $\mathcal{L}$ satisfying certain axioms, then the agent chooses as if she has a probability distribution p over Ω and a utility function U over X, and maximizes expected utility.

This theory makes no comment as to what is contained in the set X over which utilities are defined (cf. James C. Cox and Vjollca Sadiraj 2002). The argument that Alice's risk aversion over small stakes implies implausible behavior over large stakes implicitly assumes that utility is a function of (only) Alice's final wealth—the amount of money she has after the outcome of the lottery has been realized. Hence, Alice must view winning a million-dollar lottery when initially penniless as equivalent to losing \$9,000,000 of an initial \$10,000,000. This is the most common way that expected utility appears in theoretical models, but nothing in expected utility theory precludes defining utility over pairs of the form (w, y), where w is an initial wealth level and y is a gain or loss by which this wealth level is adjusted. In this case, Alice may view the two final \$1,000,000 outcomes described above quite differently. And once this is the case, there need no longer be any conflict between being an expected utility maximizer, being risk averse over small stakes, and still behaving plausibly over larger states.[65]

This argument can be taken a step further. Savage (1972, pp. 15–16, 82–91) views expected utility theory as applicable only to "small-worlds" problems, in which the sets of states, consequences and acts are simple enough that one can identify and explore every implication of each act. Savage notes that it is "utterly ridiculous" to encompass *all* of our decision-making within a single small-worlds model (1972, p. 16). Instead, his view (1972, pp. 82–91) is that decision makers break the world they face into small chunks that are simple enough to be approximated with a small-worlds view. We can expect behavior in these subproblems to be described by expected utility theory, but the theory tells us nothing about relationships between behavior across problems.

Duncan Luce and Howard Raiffa (1957, pp. 299–300) continue this argument, noting that, "one's choices for a series of problems—no matter how simple—usually are not consistent." They suggest that if one discovers an inconsistency, one should modify one's decisions, with "this jockeying—making snap judgments, checking on their consistency, modifying them, again checking on consistency, etc.", ultimately leading to consistent expected-utility maximizing behavior. We can thus expect consistent behavior only across sets of choices (or worlds) that are sufficiently small that we can expect the required adjustment to have been made.

Returning to our original setting, the set of all lotteries may be too large a world to encompass within a single expected-utility

[65] There is then no inconsistency in believing that experimental subjects are expected utility maximizers while using lotteries to control for risk aversion over small stakes. The evidence on whether lottery payoffs successfully control for risk aversion is not entirely encouraging (e.g., Joyce E. Berg, John W. Dickhaut, and Thomas A. Rietz 2003; James C. Cox and Ronald L. Oaxaca 1995; Selten, Abdolkarim Sadrieh, and Klaus Abbink 1999; and James M. Walker, Smith, and Cox 1990). These findings present yet another challenge to the presumption that experimental subjects maximize expected utility.

formulation. If we define utility in terms of final wealth levels, Alice's expected-utility maximization over small stakes may then not be consistent with her behavior over large stakes. But she may nonetheless be maximizing expected utility, though with a utility function in which wealth or some other variable indexes different small worlds problems, each of which is treated via a utility function over (some subset of) final wealth levels.

This discussion is *not* to be read as a defense of expected utility theory. There is every reason to believe that so stark a theory cannot always be a good approximation. This discussion is instead meant to provide a word of caution in assessing the internal validity of experimental results. Risk aversion over small gambles, one of the seemingly most powerful challenges to the theory, may in fact be consistent with expected utility.

More importantly, this argument does not diminish the strength of the small-stakes-risk-aversion challenge to economic theory. The evidence remains that we can save expected utility maximization as a useful theory only if something other than wealth enters utility functions. As Rubinstein (2001) notes, this opens the door to all manner of inconsistencies in decision making. Expected utility can be defended only by recognizing that economic theorists have a great deal of work to do.

Other illustrations of the importance of theory in assessing internal validity are easily found. Game-theoretic models featuring mixed Nash equilibria have been questioned on the grounds that individual play does not exhibit the identical, independent randomization required by the theory (e.g., James N. Brown and Robert W. Rosenthal 1990). But if the mixed equilibrium reflects either a population polymorphism (as suggested by John F. Nash 2002) or the result of an adaptive process, we would expect such independence to fail (e.g., Binmore, Joe Swierzbinski, and Chris Proulx 2001).

Alternatively, section 2 sketched the discussion of behavior in bargaining games up to the appearance of models in which preferences depend upon the vector of all payoffs, one's own as well as the payoffs of others. Subsequent experiments have suggested that more is involved. Attitudes towards payoffs appear to depend not only on the payoffs themselves, but also the context in which these payoffs were generated. A player is more likely to prefer a larger opponent payoff if the opponent's play has been appropriate (kind, or fair, or generous, or expected) and more likely to prefer a smaller opponent payoff if the opponent's play has been inappropriate. The experimental evidence has provided evidence for positive reciprocity (the desire to reward those who have behaved appropriately) (Fehr and Simon Gächter 2000; Fehr, Gächter, and Georg Kirchsteiger 1997; Kevin A. McCabe, Rassenti, and Smith 1998) and negative reciprocity (the desire to punish those who have behaved inappropriately) (Fehr and Gächter 2002). However, we can expect subjects' choices to reflect a mixture of concerns for one's own payoff, inequality aversion, altruism, trust, and positive and negative reciprocity (cf. Cox 2004). How do we separate these forces, i.e., how do we assess the internal validity of the experiments? Once again, a useful point of departure is a model of preferences encompassing these forces and pointing to experiments that will distinguish them. Interpreting the experiments is again likely to rest upon careful theoretical modeling.

5. *The Search for Theory*

Where do we look for theoretical developments that will help integrate economic theory and experimental economics?

To approach this question, think of an experiment as being composed of three pieces. The game form (recognizing that the "game" may include only a single-player) specifies the rules of play, including the number and characteristics of the players,

the choices available to the players, their timing and sequence, the information available to the players, the resulting consequences, and so on. To this, one adds a specification of how the outcomes are translated into utilities. It would typically be convenient if the monetary payoffs given by the game form could also be taken to represent players' utilities, but this need not be the case. Let us refer to a game form and its associated utilities as a *game protocol* or simply protocol, and let us think of the "default" protocol as equating monetary payoffs and utilities.[66] The third piece of the triad is a theory describing the behavior one would expect, given the game protocol.

In some cases, the game protocol leaves little to the discretion of the theory. If the protocol combines the dictator game with the assumption that monetary payoffs are equivalent to utilities, then a theory based on rational behavior leaves no room for maneuver: dictators must keep all of the money. Similarly, if the protocol pairs the ultimatum game with the assumption that monetary payoffs are utilities, then sequential rationality uniquely determines the implications of the theory. In other cases, the protocol leaves much to the discretion of the theory. Work on equilibrium refinements grew out of the fact that even if one restricts attention to relatively simple games and assumes that the payoffs are indeed utilities, sequential rationality in general puts relatively few restrictions on behavior.

Now consider how one might react if an economic theory and experimental results are consistently at odds. One possibility is that the theory should be refined, or extended, or altered, or abandoned. For example, the equilibrium refinements literature culminated in models of equilibrium selection centered around notions of forward induction. The experimental evidence has not been particularly supportive of forward induction,[67] suggesting that theories based on forward induction could well be reconsidered.

In other cases, there is little to be gained by looking for alternative theories while maintaining the game protocol. In the bargaining games in section 2, for example, there appears to be no way to account for the observed behavior while clinging to a model based on rationality and the default protocol. The result, as we have seen, has been a flurry of work developing alternative models of preferences.[68]

We return to the modeling of preferences in section 5.2. First, however, section 5.1 considers another possible response when examining a protocol. There may be good reasons to question whether the game form perceived by the subjects matches that embedded in the experimental game protocol.

5.1 *Perceived Protocols*

How could subjects help but perceive the proper game form? The potential behavior in an experiment is typically tightly controlled, including quite precise rules for who gets to make what choices at what times. As noted in section 3.2, a great advantage of experiments is the ability to control these details. In assessing the effects of these controls, however, we return to the idea that people, including experimental subjects, use models to make decisions.

Just as economists are forced to rely on models in their analysis, so can we expect people to rely on models when making their decisions. Given the many choices people

[66] Jörgen W. Weibull (2004) introduces the concept of a game protocol, though drawing a somewhat different line between the game form and game protocol.

[67] See, for example, Dieter Balkenborg (1994); Jordi Brandts and Holt (1992, 1993, 1995); and Cooper, Susan Garvin, and Kagel (1997, 1997).

[68] Weibull (2004) stresses the possibility that an experiment's monetary payoffs may not capture subjects' preferences and discusses the resulting difficulties involved in drawing inferences from experiments. Roth (1991) notes that, given the difficulty in controlling every aspect of subjects' preferences and expectations, it is hard to know precisely what game is involved in an experimental study.

have to make in their everyday lives, most without the time and resources that economists devote to a problem, we cannot expect people to make use of all of the information in their environment.[69] Instead, most aspects of most decisions are ignored because they are not important enough to bother with. In essence, people use models, stripping away unimportant considerations to focus on more important ones.

Similarly, we should expect experimental subjects to respond to the novelty of an experimental setting by modeling its key features. This need to rely on models when analyzing the real world ensures that researchers and experimental subjects both introduce a subjective element into their perceptions.[70] Using the notation developed above, an experiment is designed to fix an experimental design x^N. The experiment itself, however, is a situation x^∞ with the property that $x^\infty(n)=x^N(n)$ for $n=1,\ldots, N$. The choice of the aspects of the situation to bring within the experimental design, captured by N, represents the experimenter's model of the situation. Suppose that an experimental subject, confronted with the situation x^∞, similarly constructs a model. This model is itself a choice of finitely many dimensions of the infinitely-dimensioned x^∞ to take into consideration. Is there any reason to expect the subject's model to coincide with the experimenter's, i.e., to expect the subject to hit upon the same choice of salient information as did the experimenter?

We may often be able to expect the subject to come close. The experiment is typically designed so as to focus attention on x^N. However, it would be surprising if the two models matched exactly. We thus run the risk that subjects may ignore aspects of the situation that the experimenter deems critical or that the subjects may introduce aspects that the experimenter deems irrelevant.[71]

An illustration is provided by Douglas Dyer and John H. Kagel (1996). Their research is motivated by the observation that experimental subjects frequently bid too aggressively in common-value auctions. Even subjects who are experienced, professional bidders in auctions for construction contracts fall prey to the winner's curse in laboratory experiments.

Dyer and Kagel note that the auctions in which the professionals routinely bid contain some potentially important features that did not appear in the laboratory experiments. For example, the real-world auctions typically allowed bidders to withdraw winning bids, without cost, when these bids contain mistakes that are formally characterized as "arithmetic errors" but in practice are allowed to cover virtually any request to withdraw a bid (on the principal that one does not want a contractor who does not want the job). As a result, the winner can withdraw a bid that is revealed (by comparison with other bids) to be too optimistic, providing some protection against the winner's curse. It appears as if the bidders have developed rules of behavior that are effective in the context with which they are familiar, though perhaps without completely identifying the key features of the environment that make these rules work well. In bringing the resulting rules into the experiment, the subjects are reacting to a perceived protocol that appears to be familiar, but with results that appear to be anomalous when held to the standard of the protocol chosen by the experimenter.[72]

[69] How long would it take to get through the grocery store if every detail of every purchase were analyzed?

[70] Section 3.1 touched on the question of whether there is an objective reality (cf. note 11). The point here is that, regardless of whether there is, models of this reality are subjective.

[71] Psychologists frequently run experiments based on the premise (often with the help of some deception) that the experimenter and subject will perceive *different* protocols.

[72] This raises the question of when we can expect lessons learned in one context to transfer to other contexts. Such transfer will presumably be more effective the greater is the extent to which people learn not only which behavior works well, but also the reasons why the behavior works well. Cooper and Kagel (2003) provide an introduction to work on generalizing learning across contexts.

The general principle is that, just as subjects in an experiment may face effective payoffs that differ from those of the game protocol, so might they effectively play a different game. Our interpretation of experimental results can then depend importantly on how we imagine subjects perceive the game.[73]

A hypothetical illustration will be helpful. Suppose that biologists were interested in a theory that female birds preferred males with long tails, and that they did so rationally because long tails were a signal of other characteristics that make a mate particularly desirable.[74] To test this theory, an experiment is designed in which some males have plastic feathers glued to their tails.[75] Suppose that females indeed flock to the males with now strikingly long tails. How do we interpret the results? A biologist is likely to claim that the experimental results provide support for the theory. However, one can well imagine an economist claiming just the opposite, that the theory has been demonstrated to be nonsense. After all, the theory is founded on the presumption of rational behavior. This seems obviously inconsistent with exhibiting a preference for males with plastic tails, since the latter cannot be associated with the characteristics that make males desirable mates.

These differing conclusions are grounded in different assumptions about how the subjects perceive the experiment. The biologist assumes that the subject will not perceive the difference between a real tail and a plastic one or, in our terms, that the subject's model of the experiment does not accommodate plastic tails. The typical assumption in economic contexts is that, provided the experiment is sufficiently transparent and effectively presented, the subjects' model of the experiment matches the experimental design.

The example of plastic tails may seem a bit removed from human experiments. Suppose instead that the hypothesis in question is that human males are attracted to females with "hourglass" figures.[76] An experimenter tests this by showing males a variety of pornographic pictures, checking which females prompt the most enthusiastic reaction. Many males are responsive to pornography, and many will be especially responsive to females with the appropriate figure. A biologist or psychologist is again likely to interpret the experimental results as support for the theory. Once more, however, one can imagine economists interpreting the results as another blow to the contention that people (or at least males) behave rationally. How can they be rational if they react the same way to fictitious females as to real ones?

An alternative interpretation is that people behave rationally, but that they use models that do not incorporate a distinction that the experimenter takes for granted. Just as birds may have a model of the world that makes no provision for plastic tails, so may people have models that do not distinguish *perfectly* between real and fictitious females (though obviously also not treating the two identically). Why might people persist in using such models? We turn to this question in section 5.2. Before doing so, one more

[73] Uri Gneezy and Aldo Rustichini (2000, 2000) report experiments showing that behavior may be counterintuitively nonmonotonic in the scale of monetary payments, with (for example) the incidence of late pickups at a day-care center increasing as the cost increases from zero to a small amount, but then being deterred by higher costs. Their interpretation is that the increase from zero to a positive cost causes agents to think about the interaction differently. In our terms, attention has been focused on different aspects of the situation, triggering the use of a different analytical model.

[74] There is a rich body of work in biology on plant and animal signaling. See Alan Grafen (1990a, 1990b) and Rufus A. Johnstone and Grafen (1992) for theoretical models; H. C. J. Godfray and Johnstone (2000) and Johnstone (1998) for surveys; and Johnstone (1995) for an examination of the evidence.

[75] See Malte Andersson (1982), J. Hoglund, M. Eriksson, and L. E. Lindell (1990), and Anders Møller (1988) for examples of similar experiments.

[76] Again, there may be a biological basis for such tastes. See, for example, Bobbi S. Low, R. D. Alexander, and K. M. Noonan (1987) and Matt Ridley (1993).

example is useful, pushing the setting yet closer to traditional economic experiments.

Return to the point of departure for section 2, the ultimatum game. A key component of the game form is that the proposer and responder will have no subsequent interactions. Experimenters have gone to great lengths to ensure that subjects understand this, including most notably ensuring that the subjects interact anonymously. But can we preclude the possibility that subjects model the situation as if there is some possibility of future interaction? If not, then the observed behavior might be consistent with rationality, without requiring any modification in how we model preferences.[77]

Fehr and Joseph Henrich (2003) shed some light on this "phantom future" explanation, pointing to experimental studies comparing behavior with and without the prospect of future interaction. For example, experimental behavior in one-shot and repeated games is markedly different (e.g., Fehr and Gächter 2000 and Gächter and Armin Falk 2002). As Fehr and Henrich note, this provides evidence that the inability to correctly account for the future is *not* a plausible explanation for the observed behavior. These results help fill in one piece of the puzzle, but leave some more to be explored. The experiments strongly suggest that people do not treat every situation as if it has the same prospects for subsequent interaction. It then remains to ask whether subjects may still model each situation as if there is *some* prospect of future interaction, perhaps on the strength of some reasoning to the effect that one can never absolutely preclude any possibility, while still recognizing that games with an explicit future are different than games without. The idea behind this middle ground is that people may not perfectly model one-shot interactions, while still recognizing and acting on the fact that the likelihood of future interaction is quite different in different situations (just as males may recognize that they are not dealing with real females when consuming pornography, and yet have a reaction to the latter shaped by their reaction to real females).

At this point, this possibility is a hypothesis awaiting further exploration and experimentation. Before expecting too much such experimentation, however, we must ask for some theoretical guidance on how people model the situations they face. What behavior could we observe that would bolster our belief in such an explanation, or that would call it into question? In essence, we need a theory of how people use theories in shaping their behavior. This is a relatively new but important direction for economic theory.[78]

The implication is that models of subjects' perceptions of experimental game forms should take their place alongside models of preferences in explaining behavior. We see evidence for an important role in the way subjects perceive experimental protocols in the importance of framing effects.[79] Why do seemingly innocuous differences in the description of a protocol make such a difference? Presumably because they prompt subjects to use different models in analyzing the experiment.

This perspective suggests that some caution is called for when working with especially complicated experiments, not simply because it may tax the abilities of the subjects (as stressed by Binmore 1999), but also because it may expand the range of models that subjects apply to the experiment. At the same time, Harrison and John A. List (2004) caution that the context-free framing of many experiments, designed to eliminate potentially confounding factors, may instead simply invite subjects to impose their own

[77] Returning to the question raised in section 2, what do subjects have to learn in the ultimatum game? Perhaps that its futureless nature distinguishes it from other, more familiar situations, and accordingly calls for different behavior (and that enough others have also learned this).

[78] Samuelson (2001) provides one example, in which the use of models makes an explicit appearance. Also see Philippe Jehiel (2004).

[79] See Robyn M. Dawes (1988) for an early discussion of framing effects.

context.[80] Despite an experimenter's best efforts to ensure that subjects understand what they are dealing with, including careful presentations, questions, and preliminary quizzes, it is not clear when we can be confident that the subjects' models match the experimenter's.[81]

In many cases, it will be difficult to distinguish whether unexpected behavior is rooted in subjects' payoffs or their perception of the game form. A tendency to analyze the ultimatum game with a model that implicitly builds in a future may give results that look as if an agent has a preference for fairness or an antipathy for asymmetric solutions. This may be more than a coincidence. As argued in Samuelson (2004), a likely response by Nature to limitations on our reasoning ability, the same sort of limitations that prompt our use of models, may be to compensate by building arguments into our preferences that we would not expect to find when agents are perfectly rational. Hence, the two arguments are likely to be complementary rather than contradictory. Then how do we choose between them, or what use is there in considering both? These questions again suggest a quest for richer theoretical models.

5.2 *Evolutionary Foundations*

One difficulty in modeling preferences is that once we move beyond a narrow conception of self-interest, there appear to be few restrictions on the features we can attribute to preferences, and hence the behavior we can explain (cf. Andrew Postlewaite 1998). Where do we find the discipline to ensure that our models are meaningful? This danger seems all the more real once we open the door to the possibility that subjects may form their own models of the experimental situation. Where do we look for a theory of peoples' models of the world?

This section suggests an evolutionary approach to both questions. The idea is to view evolution as the biological process by which humans came to their modern form.[82] This modern form includes a host of physical characteristics—our size, our relative lack of hair, our ability to walk upright— and behavioral characteristics—our diet—that we readily attribute to the forces of evolution. We can also expect our preferences and our decision-making to have been the products of evolution.[83] The result is a "reverse engineering" approach to studying decision making. Can we plausibly make a case that a given specification of preferences, or rules for how situations are modeled and translated into decisions, might have evolved as part of a solution to an evolutionary design problem? The more easily one finds such evolutionary foundations, the more seriously should we be inclined to take the model in question.[84]

Evolutionary research abounds in maladaption stories.[85] One first identifies a

[80] Camerer and Keith Weigelt (1988) conduct an early experimental analysis of reputation models. Their results exhibit many features of reputation equilibria, but their subjects also appear to have a "homemade prior" about the information structure that is at odds with a strict interpretation of the experimental environment. It appears as if the subjects have provided a context for the experiment.

[81] In the context of the hourglass-figure experiment discussed above, one can imagine an experimenter adding a variety of additional controls to ensure that the subjects understand that pornographic females are not real, perhaps stressing this in the instructions and quizzing the subjects on the difference. But this is news to no one, and is unlikely to eliminate the effect.

[82] This distinguishes this exercise from the bulk of what has come to be known as "evolutionary game theory" or "evolutionary economics." These latter bodies of (quite diverse) work share the guiding principle that instead of optimizing, people reach decisions and markets reach outcomes through an adaptive process involving varying degrees of learning, experimentation, and trial-and-error. "Evolution" is a metaphor for this adaptive process.

[83] This view is familiar in evolutionary psychology, (e.g., Leda Cosmides and John Tooby 1992), and has ample precedent in economics (e.g., Arthur J. Robson 1992, 1996, 1996, 2001a, 2001b).

[84] Just as economists are adept at building models, evolutionary psychologists have been criticized for seemingly being able to rationalize any behavior with an evolutionary model. Steven Jay Gould and Richard C. Lewontin (1979) find these models sufficiently unpersuasive as to be deemed "just-so stories." If the evolutionary approach is to be successful, it must do more than provide such stories.

[85] Terry Burnham and Jay Phelan (2000) contains a wealth of examples.

behavior that is likely to have been an optimal response to the environment in which we evolved. One then notes that our modern environment is quite different, causing the behavior to now be quite surprising, if not counterproductive.[86] In contrast, our concern here is with behavior that evolution has designed as an *optimal* response to our environment, recognizing that this is an environment in which we must rely on preferences and on models in making our decisions.

There is a growing body of work on how our evolutionary background may have shaped our preferences. Perhaps best developed is the link between reproduction and risk taking, and hence the implications for attitudes toward risk (Arthur J. Robson 1992, 1996). Much of this material is nicely covered in Robson (2001a).

More recently, experimental evidence has mounted that people will incur costs not only to bestow benefits on others, but also to penalize others, with the preference for reward or punishment hinging upon perceptions of whether the recipient has acted appropriately or inimically. What might be the evolutionary origins of such "prosocial" behavior? Henrich (2004) offers an explanation based on cultural group selection.[87] An advantage of this model is that it provides the type of discipline required for further investigation. For example, the model suggests that we should expect multiple cultural equilibria, and hence considerable variation across cultures in the tendency to bestow benefits and costs on others. Second, a propensity to imitate or conform to the behavior of others plays an important role in the model, suggesting we look for a link between such behavior and prosocial behavior. Third, evolution is viewed as facing information constraints, so that we must in turn view preferences as tools for maximizing fitness while economizing on information. These features may be consistent with a variety of other models, and so they cannot be the end of the quest, but they provide a useful point of departure.[88]

Work on how evolution has shaped the way people model their environment is in an even earlier stage. Three illustrations will be useful.

First, the Wason selection test (1966) is now a standard example in evolutionary psychology. Experimental subjects are surprisingly prone to errors in evaluating abstract conditional statements.[89] But if asked to evaluate conditional statements posed in terms of monitoring compliance with a standard of behavior, success is much higher.[90] The suggested interpretation is that our reasoning about conditional statements evolved in a setting in which monitoring behavior was particularly important.[91] This interpretation bolsters an argument of Fehr and Henrich (2003), that

[86] For a simple example, it is likely that during most of our evolutionary history, food was both in chronically short supply and could be stored only in the form of body fat. As a result, it appears likely that an evolutionarily successful strategy was to eat as much as possible whenever possible. It is then no surprise that members of modern, wealthy societies find it difficult to avoid health-threatening overeating.

[87] The model of cultural group selection avoids many of the difficulties that have made biologists skeptical of group selection arguments. Elliott Sober and David Sloan Wilson (1998) argue that group selection lies behind preferences for altruism.

[88] In connection with the first feature, it is intriguing that the study of bargaining behavior in fifteen small-scale societies by Henrich, Boyd, Bowles, Camerer, Fehr, Gintis, and McElreath (2001) finds significant behavioral variation.

[89] For example, if given a collection of cards with numbers on one side and letters on the other, along with the hypothesis that any card with a 3 on one side has a *B* on the other, and then shown four cards bearing 3, 7, *A* and *B*, only a minority correctly identify which cards must be turned over to check the hypothesis.

[90] For example, told that a coke-drinker, beer-drinker, 17-year-old and 25-year-old are seated at a table, virtually every subject knows which ones to check for compliance with a 21-year-old drinking law.

[91] See Cosmides and Tooby (1992). The ability to recognize faces (Steven Pinker 1997) is similarly interpreted as an evolutionary response to the importance of monitoring others' behavior.

laboratory behavior exhibiting reciprocity should not be interpreted as an evolutionary maladaption.[92] At the same time, however, it raises the possibility that subjects' perceived protocols may not correctly capture incentives if their presentation is sufficiently unfamiliar.

Second, a variety of evidence suggests that people are not very good in dealing with probabilities. However, there is also evidence that people fare much better when probabilities are presented in terms of frequencies.[93] This may indicate that we spent much of our history with a frequentist view of the world. This is consistent with the possibility that people are approximately expected utility maximizers, while performing quite poorly in laboratory experiments, if the latter present probabilistic information unfamiliarly.

Third, Plott (1996) presents the "discovered preference hypothesis," suggesting that rather than coming to a decision problem with fixed and well defined preferences, people respond by combining contextual information and experience with an internal search process to discover their preferences. Similar ideas appear in Dan Ariely, George Loewenstein, and Drazen Prelec's (2003) notion of "constructed" preferences, which they illustrate with a number of experiments.[94] These results initially seem to strike at the core of economic theory, calling into question the idea of stable preferences. Notice, however, that the process by which preferences are discovered or constructed sounds much like the process Luce and Raiffa (1957) describe as the foundation for expected utility maximization (cf. section 4.2.2). Rather than suggesting that we abandon expected utility theory, the experimental results again remind us that the theory may not be as straightforward as one would like. We can expect consistent behavior in settings amenable to small-worlds modeling, but must expect anomalies to appear in other situations. In terms of experiments, the implication is again that behavior may be quite sensitive to seemingly irrelevant details of the experimental environment.

What is the common theme of these three examples? Echoing the ideas that opened section 5.1, it is that evolution has equipped us with a variety of models and rules and shortcuts for dealing with a dauntingly complex world.[95] The possibility the people do not perfectly model futureless protocols may

[92] The maladaption account of such behavior would be that we evolved in an environment in which repeated or kinship interactions, and hence the optimality of reciprocity, were sufficiently pervasive that there was no point in checking whether such behavior is warranted. However, it is not clear that our evolutionary past would have equipped us with a basic propensity to monitor for and detect cheating in an environment in which it was unimportant to distinguish situations meriting reciprocity from those that do not.

[93] See Cosmides and Tooby (1996), Gerd Gigerenzer (1991, 1996, 1998), and Tversky and Kahneman (1983). For example, when told that 2 percent of the population has a disease and that a test produces no false negatives but 5 percent false positives, many subjects will struggle to ascertain the implications of a positive report. However, they fare better when told that out of every thousand people, all twenty who have the disease turn up positive but so do fifty others.

[94] For example, they find that, if subjects are first asked whether they would be willing to purchase a product at a price equal to the last two digits of their social security number, and are then asked their valuation of the product, there is significant correlation between their social security numbers and reported valuations. The suggested interpretation is that the subjects subconsciously use the numbers involved in the first purchase decision as clues to the appropriate valuation in the second. This reliance on contextual information may work well in many applications, but leads to apparently absurd behavior in the experiment.

[95] Evolutionary psychologists find evidence for constraints on evolution's ability to simply enhance our reasoning powers and dispense with these devices in the relatively large amount of energy required to maintain the human brain (Katharine Milton 1988), the high risk of maternal death in childbirth posed by infants' large heads (W. Leutenegger 1982), and the similarly-caused lengthy period of human postnatal development (Paul H. Harvey, R. D. Martin, and T. H. Clutton-Brock 1986). Andy Clark (1993) discusses the potential advantages of using contextual clues and specialized rules to conserve on generalized reasoning resources.

then arise not because evolution has erroneously designed us for a different environment, but because evolution has effectively designed us for an environment in which such shortcuts are valuable.

The first two examples are concerned with techniques for processing information. The third reflects a spillover into preferences, bringing us back to the fact that information constraints also feature in models of preference evolution. Samuelson and Jeroen Swinkels (2001) consider the relationship between these, examining the implications for preferences of an evolutionary process that must cope with scarce reasoning resources. The conclusion is that we can expect a variety of seemingly nonstandard features to be built into our preferences in response to imperfections and limitations in our information processing and reasoning.

What are the implications for economics? We can often expect people to act consistently and rationally, given their preferences. However, the preferences involved in the resulting optimizing behavior may involve all sorts of features that at first blush do not appear consistent with either the pursuit of individual self-interest or a narrow concept of consistency. Finally, we can expect context to be important. In this sense, evolution and the axiomatic approaches of Savage (1972) and Luce and Raiffa (1957) are on the same page. Both suggest that we can expect consistent behavior within sufficiently constrained contexts, though with preferences that may appear to go beyond a narrow conception of self-interest, but that the context will be important and that seeming anomalies may readily arise across contexts.

Smith (2003) argues that we can usefully view human behavior in terms of two types of rationality, "constructivist" and "ecological" rationality. Constructivist rationality resembles the rational choice models of traditional economic theory, though again allowing the possibility that preferences might reflect more than a narrow self-interest, while ecological rationality is concerned with "the possible intelligence embodied in the rules, norms, and institutions of our cultural and biological heritage...." (2003, p. 470). The argument here ties these concepts together with the vision of an evolutionary process that struggles with the constraints imposed by scarce reasoning resources. The quest to maximize fitness generates a motivation for behavior to reflect constructionist rationality, while the quest to relax constraints leads to the ready incorporation of ecological forces.

What are the implications for combining economic theory and experiments? A first one is that we must be careful in assessing both experimental findings and economic theory. For example, numerous experiments have found that subjects are willing to pay less to receive an object than they are willing to accept to relinquish the object. Some have interpreted this as reflecting a common feature of preferences, being an illustration of the more general principle that people value losses more heavily than gains.[96] However, as Plott and Kathryn Zeiler (2003) argue, there have also been many cases in which such discrepancies do not appear, and one can identify experimental settings in which the effect reliably does or does not appear. This suggests that we should stop short of proclaiming the endowment effect a universal feature of preferences, and focuses attention on the internal validity of the experiments. What links do we draw between differences in experimental settings and the forces that shape valuations? At the same time, it indicates that something is missing from our theoretical repertoire, which currently

[96] For example, Jack L. Knetsch, Fang-Fang Tang, and Thaler (2003), who also provide references to earlier work, comment that "The endowment effect and loss aversion have been among the most robust findings of the psychology of decision making. People commonly value losses much more than commensurate gains . . . " (2001, p. 257). Kahneman and Tversky (1979) stress that people view gains and losses quite differently.

provides little insight into such forces. A first step in addressing this issue would then be the construction of theoretical models, especially models shedding insight into how and when it might have been evolutionarily valuable to condition valuations on ownership.

6. *Conclusion*

Economic theory and economic experiments can be combined to the benefit of both. By itself, this is a fairly uninformative "more is better" conclusion. There must be gains from considering experiments or theory more carefully when doing theory or experimentation. But what steps can we take to make it more likely that potential gains are realized?

The danger with the concerns raised in this essay is that they might be used to apologize away any potential interaction between theory and experiments. It is unlikely that we will usefully combine theory and experiments if we too freely respond to contrasts between the two with such statements as: "The results appear to be at odds with the theory, but we have no obvious way to measure how far away they are, and by my preferred measure they are pretty close." " … but I suspect the subjects really perceived a different experimental protocol under which their behavior is consistent with the theory." " … but the theory is an approximation that cannot be expected to apply everywhere, and the discovery of this exception tells us nothing about the theory in other applications." How do we avoid working at such cross-purposes?

A good beginning would be for exercises in economic theory to routinely identify behavior that would be consistent with the theory, and especially behavior that would distinguish the theory from contending explanations. Section 3.1 noted that predicting behavior is not the only goal of economic theory, and so we cannot expect all theoretical exercises to be in a position to point to such behavior.[97] We must also allow the possibility that making connections to behavior is a goal that the theory will often not yet be sufficiently advanced to address. But at some point some connections must be made between theory and behavior if economic theory is not to fade into either philosophy or mathematics, and work that aspires to make this connection should be explicit about the implied behavior. In the course of doing so, it would be helpful to have some idea not only of the expected behavior itself, but also of how much noise we might expect to surround this behavior.

Perhaps more importantly, it would be useful for theory to identify behavior for which the theory cannot account, in the sense that the observations would force the theorist to reconsider. This would ensure that the theory is not performing well by "theorizing to the test," as in section 4.1.1. The behavior relegated to this category might further be grouped in two categories. One, recognizing that theories can be useful without applying universally, would identify situations that are not a good match for the theory and in which contrary behavior would not shake one's confidence in the usefulness of the theory. The second would consist of behavior that would force reconsideration of the theory. The strength of a theory will often be reflected in the content of this latter category, and we might move toward an explicit examination of what makes a theory useful.

Similarly, it would be helpful to have the experimental design indicate which outcomes would be regarded as a failure as well

[97] For example, the theory of utility maximization occupies a prized place at the center of economics. However, the theory has very little predictive content. Given the freedom to define preferences, virtually any behavior can be reconciled with expected-utility maximization. Even apparent violations of the axioms of revealed preference can often be apologized away by noting that the data consist of choices made at different times and thus while the decision-maker is in different states, and hence possibly described by different preferences. Predictions then require some augmenting or extension of the revealed-preference axioms, as in Cox (1997).

as which would be considered a success. This question appears to be trivial in many cases, with success and failure riding on the statistical significance of an estimated parameter. However, one of the advantages of experimental work is the ability to control the environment and design the tests. This allows us to direct attention away from issues of statistical significance and toward issues of economic importance. The strength of the experiment will often be reflected in the content of this "failure" category.

Finally, again returning to section 4.1.1, it is important that both theoretical models and interpretations of experimental results be precise enough to apply beyond the experimental situation from which they emerge. This allows links to be made that multiply the power of single studies.

7. *Appendix: Proof of Proposition 1*

Proof. Given the design x^N, the experiment induces a true probability distribution $\pi^* \in \Delta S^M$ over elements s^M of the set S^M, while the theory f induces a probability distribution $f^* \in \Delta\Delta S^M$ over elements π of the set ΔS^M. Given f^* and π^*, let

$$H(f^*,\pi^*) = \int_{S^M}\int_{\Delta S^M} T(s^M,\pi)df^*(\pi)d\pi^*(s^M).$$

$H(f^*, \pi^*)$is thus the probability the theory is accepted. Let f_{π^*} be a measure over ΔS^M that puts probability one on the true distribution π^*. Then, from the assumption that T accepts the truth with probability $1-\epsilon$, we have, for any π^*,

$$H(f_{\pi^*},\pi^*) \geq 1-\epsilon. \tag{5}$$

We need to show there exists an $\hat{f} \in \Delta\Delta S^M$ such that, for all $s^M \in S^M$

$$H(\hat{f},\pi_{s^M}) \geq 1-\epsilon, \tag{6}$$

where π_{s^M} is a probability distribution over S^M that puts unitary probability on outcome s^M. Using the definition of a maximum for the first inequality and (5) for the second, we have

$$\min_{\pi^* \in \Delta S^M} \max_{f^* \in \Delta\Delta S^M} H(f^*,\pi^*) \geq \min_{\pi^* \in \Delta S^M} H(f_{\pi^*},\pi^*) \geq 1-\epsilon. \tag{7}$$

From Ky Fan's (1953, Theorem 1) first minmax theorem, we have:

$$\min_{\pi^* \in \Delta S^M} \max_{f^* \in \Delta\Delta S^M} H(f^*,\pi^*) = \max_{f^* \in \Delta\Delta S^M} \min_{\pi^* \in \Delta S^M} H(f^*,\pi^*). \tag{8}$$

Using (8) to replace the first term in (7) and deleting the middle term then gives

$$\max_{f^* \in \Delta\Delta S^M} \min_{\pi^* \in \Delta S^M} H(f,\pi^*) \geq 1-\epsilon, \tag{9}$$

and hence, there is an $\hat{f}$ such that, for all $s^M \in S^M$,

$$H(\hat{f},\pi_{s^M}) \geq 1-\epsilon$$

which, from (6), is the desired result.

References

Anderson, Simon P., Jacob K. Goeree, and Charles A. Holt. 1998. "Rent Seeking with Bounded Rationality: An Analysis of the All-Pay Auction." *Journal of Political Economy*, 106(4): 828–53.

Anderson, Simon P., Jacob K. Goeree, and Charles A. Holt. 1998. "A Theoretical Analysis of Altruism and Decision Error in Public Goods Games." *Journal of Public Economics*, 70(2): 297–323.

Anderson, Simon P., Jacob K. Goeree, and Charles A. Holt. 2001. "Minimum-Effort Coordination Games: Stochastic Potential and Logit Equilibrium." *Games and Economic Behavior*, 34(2): 177–99.

Andersson, Malte. 1982. "Female Choice Selects for Extreme Tail Length in a Widowbird." *Nature*, 299(5886): 818–20.

Andreoni, James, Paul M. Brown, and Lise Vesterlund. 2002. "What Makes an Allocation Fair? Some Experimental Evidence." *Games and Economic Behavior*, 40(1): 1–24.

Andreoni, James, Marco Castillo, and Ragan Petrie. 2003. "What Do Bargainers' Preferences Look Like? Experiments with a Convex Ultimatum Game." *American Economic Review*, 93(3): 672–85.

Andreoni, James and John Miller. 2002. "Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism." *Econometrica*, 70(2): 737–53.

Andreoni, James and Larry Samuelson. 2003. "Building

Rational Cooperation," SSRI Working Paper 2003-4.
Ariely, Dan, George Loewenstein, and Drazen Prelec. 2003. "'Coherent Arbitrariness': Stable Demand Curves without Stable Preferences." *The Quarterly Journal of Economics*, 118(1): 73–105.
Aumann, Robert J. 1985. "What Is Game Theory Trying to Accomplish?" in *Frontiers of Economics*. Kenneth J. Arrow and Seppo Honkapohja, eds. Oxford: Blackwell, 28–76.
Balkenborg, Dieter. 1994. "An Experiment on Forward versus Backward Induction," SFB Discussion Paper B-268, University of Bonn.
Banks, Jeffrey S., John O. Ledyard, and David P. Porter. 1989. "Allocating Uncertain and Unresponsive Resources: An Experimental Approach." *Rand Journal of Economics*, 20(1): 1–25.
Bateson, Melissa and Alex Kacelnik. 1996. "Rate Currencies and the Foraging Starling: The Fallacy of the Averages Revisited." *Behavioral Ecology*, 7(3): 341–52.
Battalio, Raymond, Larry Samuelson, and John Van Huyck. 2001. "Optimization Incentives and Coordination Failure in Laboratory Stag Hunt Games." *Econometrica*, 69(3): 749–64.
Berg, Joyce E., John W. Dickhaut, and Thomas A. Rietz. 2003. "Diminishing Preference Reversals by Inducing Risk Preferences: Evidence for Noisy Maximization." *Journal of Risk and Uncertainty*, 27(2): 139–70.
Bergstrom, Theodore C. 2003. "Vernon Smith's Insomnia and the Dawn of Economics As Experimental Science." *Scandinavian Journal of Economics*, 105(2): 181–205.
Binmore, Ken. 1999. "Why Experiment in Economics?" *Economic Journal*, 109(453): F16–24.
Binmore, Ken and Paul Klemperer. 2002. "The Biggest Auction Ever: The Sale of the British 3G Telecom Licenses." *Economic Journal*, 112(478): C74–96.
Binmore, Ken, John McCarthy, Giovanni Ponti, Larry Samuelson, and Avner Shaked. 2002. "A Backward Induction Experiment." *Journal of Economic Theory*, 104(1): 48–88.
Binmore, Ken, Peter Morgan, Avner Shaked, and John Sutton. 1991. "Do People Exploit Their Bargaining Power? An Experimental Study." *Games and Economic Behavior*, 3(3): 295–322.
Binmore, Ken and Larry Samuelson. 1999. "Evolutionary Drift and Equilibrium Selection." *The Review of Economic Studies*, 66(2): 363–93.
Binmore, Ken, Avner Shaked, and John Sutton. 1985. "Testing Noncooperative Bargaining Theory: A Preliminary Study." *American Economic Review*, 75(5): 1178–80.
Binmore, Ken, Avner Shaked, and John Sutton. 1989. "An Outside Option Experiment." *The Quarterly Journal of Economics*, 104(4): 753–70.
Binmore, Ken, Joe Swierzbinski, and Chris Proulx. 2001. "Does Minimax Work? An Experimental Study." *Economic Journal*, 111(473): 445–64.
Binmore, Kenneth G., John Gale, and Larry Samuelson. 1995. "Learning to Be Imperfect: The Ultimatum Game." *Games and Economic Behavior*, 8(1): 56–90.
Blount, Sally. 1995. "When Social Outcomes Aren't Fair: The Effect of Causal Attributions on Preferences." *Organizational Behavior and Human Decision Processes*, 63(2): 131–44.
Bolton, Gary E. 1991. "A Comparative Model of Bargaining: Theory and Evidence." *American Economic Review*, 81(5): 1096–136.
Bolton, Gary E. and Axel Ockenfels. 2000. "ERC: A Theory of Equity, Reciprocity, and Competition." *American Economic Review*, 90(1): 166–93.
Bolton, Gary E. and Rami Zwick. 1995. "Anonymity versus Punishment in Ultimatum Bargaining." *Games and Economic Behavior*, 10(1): 95–121.
Brandts, Jordi and Charles A. Holt. 1992. "An Experimental Test of Equilibrium Dominance in Signaling Games." *American Economic Review*, 82(5): 1350–65.
Brandts, Jordi and Charles A. Holt. 1993. "Adjustment Patterns and Equilibrium Selection in Experimental Signaling Games." *International Journal of Game Theory*, 22(3): 279–302.
Brandts, Jordi and Charles A. Holt. 1995. "Limitations of Dominance and Forward Induction: Experimental Evidence." *Economics Letters*, 49(4): 391–95.
Brewer, Paul J., Maria Huang, Brad Nelson, and Charles R. Plott. 2002. "On the Behavioral Foundations of the Law of Supply and Demand: Human Convergence and Robot Randomness." *Experimental Economics*, 5(3): 179–208.
Brewer, Paul J. and Charles R. Plott. 1996. "A Binary Conflict Ascending Price (BICAP) Mechanism for the Decentralized Allocation of the Right to Use Railroad Tracks." *International Journal of Industrial Organization*, 14(6): 857–86.
Brown, James N. and Robert W. Rosenthal. 1990. "Testing the Minimax Hypothesis: A Re-examination of O'Neill's Game Experiment." *Econometrica*, 58(5): 1065–81.
Bulmer, Michael. 1997. *Theoretical Evolutionary Ecology*. Sunderland, MA: Sinauer Associates, Inc.
Burnham, Terry and Jay Phelan. 2000. *Mean Genes: From Sex to Money to Food: Taming Our Primal Instincts*. Cambridge: Perseus Publishing.
Camerer, Colin. 1995. "Individual Decision Making," in *Handbook of Experimental Economics*. John H. Kagel and Alvin E. Roth, eds. Princeton: Princeton University Press: 587–703.
Camerer, Colin. 2003. *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton: Princeton University Press and Russell Sage Foundation.
Camerer, Colin and Keith Weigelt. 1988. "Experimental Tests of a Sequential Equilibrium Reputation Model." *Econometrica*, 56(1): 1–36.
Cameron, Lisa A. 1999. "Raising the Stakes in the Ultimatum Game: Experimental Evidence from Indonesia." *Economic Inquiry*, 37(1): 47–59.
Cason, Timothy N. and Daniel Friedman. 1996. "Price Formation in Double Auction Markets." *Journal of Economic Dynamics and Control*, 20(8): 1307–37.
Charness, Gary and Matthew Rabin. 2002. "Understanding Social Preferences with Simple Tests." *The Quarterly Journal of Economics*, 117(3): 817–69.
Clark, Andy. 1993. *Microcognition: Philosophy,*

Cognitive Science and Parallel Distributed Processing. Cambridge: MIT Press.

Coller, Maribeth, Glenn W. Harrison, and E. Elisabet Rutström. 2003. "Are Discount Rates Constant? Reconciling Theory and Observation." Mimeo. University of South Carolina and University of Central Florida.

Cooper, David J., Nick Feltovich, Alvin E. Roth, and Rami Zwick. 2003. "Relative versus Absolute Speed of Adjustment in Strategic Environments: Responder Behavior in Ultimatum Games." *Experimental Economics*, 6(2): 181–207.

Cooper, David J., Susan Garvin, and John H. Kagel. 1997. "Adaptive Learning vs. Equilibrium Refinements in an Entry Limit Pricing Game." *Economic Journal*, 107(442): 553–75.

Cooper, David J., Susan Garvin, and John H. Kagel. 1997. "Signalling and Adaptive Learning in an Entry Limit Pricing Game." *Rand Journal of Economics*, 28(4): 662–83.

Cooper, David J. and John H. Kagel. 2003. "Lessons Learned: Generalizing Learning across Games." *American Economic Review*, 93(2): 202–07.

Cosmides, Leda and John Tooby. 1992. "The Psychological Foundations of Culture," in *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. Jerome H. Barkow, Leda Cosmides, and John Tooby, eds. Oxford: Oxford University Press, 19–136.

Cosmides, Leda and John Tooby. 1996. "Are Humans Good Intuitive Statisticians After All? Rethinking Some Conclusions from the Literature on Judgement Under Uncertainty." *Cognition*, 58(1): 1–73.

Cosmides, Leda and John Tooby. 1996. "Cognitive Adaptations for Social Exchange," in *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. Jerome H. Barkow, Leda Cosmides, and John Tooby, eds. Oxford: Oxford University Press: 163–228.

Cox, James C. 1997. "On Testing the Utility Hypothesis." *Economic Journal*, 107(443): 1054–78.

Cox, James C. 2004. "How to Identify Trust and Reciprocity." *Games and Economic Behavior*, 46(2): 260–81.

Cox, James C. and Ronald L. Oaxaca. 1995. "Inducing Risk-Neutral Preferences: Further Analysis of the Data." *Journal of Risk and Uncertainty*, 11(1): 65–79.

Cox, James C. and Vjollca Sadiraj. 2002. "Risk Aversion and Expected Utility Theory: Coherence for Small- and Large-Stakes Gambles." Mimeo. University of Arizona and University of Amsterdam.

Crawford, Vincent P. 1997. "Theory and Experiment in the Analysis of Strategic Interaction," in *Advances in Economics and Econometrics: Theory and Applications, Volume 1*. Davis M. Kreps and Kenneth F. Wallis, eds. Cambridge: Cambridge University Press, 206–42.

Dasgupta, Partha and Eric Maskin. 2003. "Uncertainty, Waiting Costs, and Hyperbolic Discounting." Mimeo. Institute for Advanced Study, Princeton.

Davis, Douglas D. and Charles A. Holt. 1993. *Experimental Economics*. Princeton: Princeton University Press.

Dawes, Robyn M. 1988. *Rational Choice in an Uncertain World*. New York: Harcourt Brace Jovanovich.

Dekel, Eddie and Yossi Feinberg. 2004. "A True Expert Knows Which Questions Should be Asked," Mimeo. Institute for Advanced Study, Princeton.

Dennett, Daniel C. 1995. *Darwin's Dangerous Idea*. New York: Simon and Schuster.

Drago, Robert and John S. Heywood. 1989. "Tournaments, Piece Rates, and the Shape of the Payoff Function." *Journal of Political Economy*, 97(4): 992–98.

Dufwenberg, Martin and Georg Kirchsteiger. 2004. "A Theory of Sequential Reciprocity." *Games and Economic Behavior*, 47(2): 268–98.

Dyer, Douglas and John H. Kagel. 1996. "Bidding in Common Value Auctions: How the Commercial Construction Industry Corrects for the Winner's Curse." *Management Science*, 42(10): 1463–75.

El-Gamal, Mahmoud A. and David M. Grether. 1995. "Are People Bayesian? Uncovering Behavioral Strategies." *Journal of the American Statistical Association*, 90(432): 1137–45.

Erev, Ido, Alvin E. Roth, Robert L. Slonim, and Greg Barron. 2002. "Combining a Theoretical Prediction With Experimental Evidence." Mimeo. Harvard University.

Fagin, Ronald, Joseph Y. Halpern, Yoram Moses, and Moshe Y. Vardi. 1995. *Reasoning About Knowledge*. Cambridge: MIT Press.

Falk, Armin, Ernst Fehr, and Urs Fischbacher. 2003. "On the Nature of Fair Behavior." *Economic Inquiry*, 41(1): 20–26.

Fan, Ky. 1953. "Minimax Theorems." *Proceedings of the National Academy of Sciences of the United States of America*, 39: 42–47.

Fehr, Ernst and Simon Gächter. 2000. "Cooperation and Punishment in Public Goods Experiments." *American Economic Review*, 90(4): 980–94.

Fehr, Ernst and Simon Gächter. 2000. "Fairness and Retaliation: The Economics of Reciprocity." *Journal of Economic Perspectives*, 14(3): 159–81.

Fehr, Ernst and Simon Gächter. 2002. "Altruistic Punishment in Humans." *Nature*, 415(6868): 137–40.

Fehr, Ernst, Simon Gächter, and Georg Kirchsteiger. 1997. "Reciprocity as a Contract Enforcement Device: Experimental Evidence." *Econometrica*, 65(4): 833–60.

Fehr, Ernst and Joseph Henrich. 2003. "Is Strong Reciprocity a Maladaptation? On the Evolutionary Foundations of Human Altruism," in *The Genetic and Cultural Evolution of Cooperation*. Peter Hammerstein, ed. Cambridge: MIT Press: 55–82.

Fehr, Ernst and Klaus M. Schmidt. 1999. "A Theory of Fairness, Competition, and Cooperation." *The Quarterly Journal of Economics*, 114(3): 817–68.

Forsythe, Robert, Joel L. Horowitz, N. E. Savin, and Martin Sefton. 1994. "Fairness in Simple Bargaining Experiments." *Games and Economic Behavior*, 6(3): 347–69.

Frederick, Shane, George Loewenstein, and Ted O'Donoghue. 2002. "Time Discounting and Time Preference: A Critical Review." *Journal of Economic*

Literature, 40(2): 351–401.
Fudenberg, Drew and David K. Levine. 1997. "Measuring Players' Losses in Experimental Games." *The Quarterly Journal of Economics*, 112(2): 507–36.
Fudenberg, Drew and David K. Levine. 1998. *The Theory of Learning in Games*. Cambridge: MIT Press.
Gächter, Simon and Armin Falk. 2002. "Reputation and Reciprocity: Consequences for the Labour Relation." *Scandinavian Journal of Economics*, 104(1): 1–26.
Geanakoplos, John, David Pearce, and Ennio Stacchetti. 1989. "Psychological Games and Sequential Rationality." *Games and Economic Behavior*, 1(1): 60–79.
Gigerenzer, Gerd. 1991. "How to Make Cognitive Illusions Disappear: Beyond Heuristics and Biases," in *European Review of Social Psychology, Volume 2*. Wolfgang Stroebe and Miles Hewstone, eds. New York: Wiley.
Gigerenzer, Gerd. 1996. "The Psychology of Good Judgment: Frequency Formats and Simple Algorithms." *Journal of Medical Decision Making*, 16(3): 273–80.
Gigerenzer, Gerd. 1998. "Ecological Intelligence: An Adaptation for Frequencies," in *The Evolution of Mind*. D. Cummings and C. Allen, eds. Oxford: Oxford University Press: 9–29.
Gneezy, Uri and Aldo Rustichini. 2000. "A Fine is a Price." *Journal of Legal Studies*, 29(1): 1–18.
Gneezy, Uri and Aldo Rustichini. 2000. "Pay Enough or Don't Pay at All." *The Quarterly Journal of Economics*, 115(3): 791–810.
Gode, Dhananjay K. and Shyam Sunder. 1993. "Allocative Efficiency of Markets with Zero-Intelligence Traders: Market as a Partial Substitute for Individual Rationality." *Journal of Political Economy*, 101(1): 119–37.
Gode, Dhananjay K. and Shyam Sunder. 1993. "Lower Bounds for Efficiency of Surplus Extraction in Double Auctions," in *The Double Auction Market: Institutions, Theories and Evidence, Proceedings Vol. 15*. Daniel Friedman and John Rust, eds. Santa Fe: Santa Fe Institute in the Science of Complexity.
Gode, Dhananjay K. and Shyam Sunder. 1997. "What Makes Markets Allocationally Efficient?" *The Quarterly Journal of Economics*, 112(2): 603–30.
Godfray, H. C. J. and R. A. Johnstone. 2000. "Begging and Bleating: The Evolution of Parent-Offspring Signalling." *Philosophical Transactions of the Royal Society of London B*, 355(1403): 1581–91.
Goeree, Jacob K. and Charles A. Holt. 2001. "Ten Little Treasures of Game Theory and Ten Intuitive Contradictions." *American Economic Review*, 91(5): 1402–22.
Goeree, Jacob K., Charles A. Holt, and Thomas R. Palfrey. 2002. "Quantal Response Equilibrium and Overbidding in Private-Value Auctions." *Journal of Economic Theory*, 104(1): 247–72.
Goeree, Jacob K., Charles A. Holt, and Thomas R. Palfrey. 2004. "Regular Quantal Response Equilibrium," Mimeo. California Institute of Technology and University of Virginia.
Gould, Steven Jay and Richard C. Lewontin. 1979. "The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptionist Programme." *Proceedings of the Royal Society of London, Series B*, 205(1161): 581–98.
Grafen, Alan. 1990a. "Biological Signals as Handicaps." *Journal of Theoretical Biology*, 144(4): 517–46.
Grafen, Alan. 1990b. "Sexual Selection Unhandicapped by the Fisher Process." *Journal of Theoretical Biology*, 144(4): 473–516.
Güth, Werner, Rolf Schmittberger, and Bernd Schwarze. 1982. "An Experimental Analysis of Ultimatum Bargaining." *Journal of Economic Behavior and Organization*, 3(4): 367–88.
Güth, Werner and Reinhard Tietz. 1990. "Ultimatum Bargaining Behavior: A Survey and Comparison of Experimental Results." *Journal of Economic Psychology*, 11(3): 417–49.
Haile, Philip, Ali Hortacsu, and Grigory Kosenok. 2004. "On the Empirical Content of Quantal Response Equilibrium," Mimeo. Yale University.
Halevy, Yoram. 2004. "Diminishing Impatience: Disentangling Time Preference from Uncertain Lifetime," Mimeo. University of British Columbia.
Harless, David W. and Colin F. Camerer. 1994. "The Predictive Utility of Generalized Expected Utility Theories." *Econometrica*, 62(6): 1251–89.
Harless, David W. and Colin F. Camerer. 1995. "An Error Rate Analysis of Experimental Data Testing Nash Refinements." *European Economic Review*, 39(3–4): 649–60.
Harrison, Glenn W. 1989. "Theory and Misbehavior of First-Price Auctions." *American Economic Review*, 79(4): 749–62.
Harrison, Glenn W. 1990. "Risk Attitudes in First-Price Auction Experiments: A Bayesian Analysis." *Review of Economics and Statistics*, 72(3): 541–46.
Harrison, Glenn W. 1992. "Theory and Misbehavior of First-Price Auctions: Reply." *American Economic Review*, 82(5): 1426–43.
Harrison, Glenn W., Ronald M. Harstad, and E. Elisabet Rutström. 2002. "Experimental Methods and Elicitation of Values," University of South Carolina Working Paper B-95-11.
Harrison, Glenn W. and Jack Hirshleifer. 1989. "An Experimental Evaluation of Weakest Link/Best Shot Models of Public Goods." *Journal of Political Economy*, 97(1): 201–25.
Harrison, Glenn W. and John A. List. 2004. "Field Experiments," Mimeo. University of Central Florida and University of Maryland.
Harrison, Glenn W. and Kevin A. McCabe. 1992. "Testing Non-Cooperative Bargaining Theory in Experiments," in *Research in Experimental Economics, Volume 5*. R. Mark Isaac, ed. London: JAI Press, 137–69.
Harrison, Glenn W. and Kevin A. McCabe. 1996. "Expectations and Fairness in a Simple Bargaining Experiment." *International Journal of Game Theory*, 25(3): 303–27.
Harvey, Paul H., R. D. Martin, and T. H. Clutton-Brock. 1986. "Life Histories in Comparative Perspective," in *Primate Societies*. Barbara B. Smuts, Dorothy L. Cheeney, Robert M. Seyfarth, Richard

W. Wrangham, and Thomas T. Struhsaker, eds. Chicago: University of Chicago Press, 181–96.

Henrich, Joseph. 2000. "Does Culture Matter in Economic Behavior? Ultimatum Game Bargaining among the Machiguenga of the Peruvian Amazon." *American Economic Review*, 90(4): 973–79.

Henrich, Joseph. 2004. "Cultural Group Selection, Coevolutionary Processes and Large-Scale Cooperation." *Journal of Economic Behavior and Organization*, 53(1): 3–35.

Henrich, Joseph, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, Herbert Gintis, and Richard McElreath. 2001. "In Search of Homo Economicus: Behavioral Experiments in 15 Small-Scale Societies." *American Economic Review*, 91(2): 73–78.

Hoffman, Elizabeth, Kevin A. McCabe, and Vernon L. Smith. 1996. "On Expectations and the Monetary Stakes in Ultimatum Games." *International Journal of Game Theory*, 25(3): 289–301.

Hoffman, Elizabeth, Kevin McCabe, Keith Shachat, and Vernon L. Smith. 1994. "Preferences, Property Rights, and Anonymity in Bargaining Games." *Games and Economic Behavior*, 7(3): 346–80.

Hoffman, Elizabeth, Kevin McCabe, and Vernon L. Smith. 1996. "Social Distance and Other-Regarding Behavior in Dictator Games." *American Economic Review*, 86(3): 653–60.

Hoglund, J., M. Eriksson, and Lindell. L. E. 1990. "Females of the Lek-Breeding Great Snipe, *Gallinago media,* Prefer Males with White Tails." *Animal Behavior*, 40: 23–32.

Holt, Charles A. and Susan K. Laury. 2002. "Risk Aversion and Incentive Effects." *American Economic Review*, 92(5): 1644–55.

Hopkins, Ed. 2002. "Two Competing Models of How People Learn in Games." *Econometrica*, 70(6): 2141–66.

Houser, Daniel, Michael Keane, and Kevin McCabe. 2004. "Behavior in a Dynamic Decision Problem: An Analysis of Experimental Evidence Using a Bayesian Type Classification Algorithm." *Econometrica*, 72(3): 781–822.

Houston, Alasdair I. and John M. McNamara. 1999. *Models of Adaptive Behavior: An Approach Based on State*. Cambridge: Cambridge University Press.

Jehiel, Philippe. 2004. "Analogy-Based Expectation Equilibrium." *Journal of Economic Theory*, In Press.

Johnstone, Rufus A. 1995. "Sexual Selection, Honest Advertisement and the Handicap Principle: Reviewing The Evidence." *Biological Reviews*, 70(1): 1–65.

Johnstone, Rufus A. 1998. "Game Theory and Communication," in *Game Theory and Animal Behavior*. Lee Alan Dugatkin and Hudson Kern Reeve, eds. Oxford: Oxford University Press, 94–117.

Johnstone, Rufus A. and Alan Grafen. 1992. "Error-Prone Signalling." *Proceedings of the Royal Society of London, Series B*, 248: 229–33.

Kacelnik, Alex. 1997. "Normative and Descriptive Models of Decision Making: Time Discounting and Risk Sensitivity," in *Characterizing Human Psychological Adaptations*. Gregory R. Bock and Gail Cardew, eds. New York: Wiley, 51–70.

Kacelnik, Alex and Fausto Brito e Abreu. 1998. "Risky Choice and Weber's Law." *Journal of Theoretical Biology*, 194(2): 289–98.

Kagel, John H. 1995. "Auctions: A Survey of Experimental Research," in *The Handbook of Experimental Economics*. John H. Kagel and Alvin E. Roth, eds. Princeton: Princeton University Press, 501–85.

Kagel, John H., Ronald M. Harstad, and Dan Levin. 1987. "Information Impact and Allocation Rules in Auctions with Affiliated Private Values: A Laboratory Study." *Econometrica*, 55(6): 1275–304.

Kagel, John H. and Dan Levin. 2002. *Common Value Auctions and the Winner's Curse*. Princeton and Oxford: Princeton University Press.

Kahneman, Daniel and Amos Tversky, eds. 2000. *Choices, Values, and Frames*. Cambridge: Cambridge University Press; New York: Russell Sage Foundation.

Kahneman, Daniel and Amos Tversky. 1979. "Prospect Theory: An Analysis of Decision under Risk." *Econometrica*, 47(2): 263–91.

Knetsch, Jack L., Fang-Fang Tang, and Richard H. Thaler. 2001. "The Endowment Effect and Repeated Market Trials: Is the Vickrey Auction Demand Revealing?" *Experimental Economics*, 4(3): 257–69.

Kohlberg, Elon and Jean-Francois Mertens. 1986. "On the Strategic Stability of Equilibria." *Econometrica*, 54(5): 1003–37.

Ledyard, John O. 1986. "The Scope of the Hypothesis of Bayesian Equilibrium." *Journal of Economic Theory*, 39(1): 59–82.

Ledyard, John O., David Porter, and Randii Wessen. 2000. "A Market-Based Mechanism for Allocating Space Shuttle Secondary Payload Priority." *Experimental Economics*, 2(3): 173–95.

Leutenegger, W. 1982. "Encephalization and Obstetrics in Primates with Particular Reference to Human Evolution," in *Primate Brain Evolution: Methods and Concepts*. Este Armstrong and Dean Falk, eds. New York: Plenum: 85–95.

Levine, David K. 1998. "Modeling Altruism and Spitefulness in Experiments." *Review of Economic Dynamics*, 1(3): 593–622.

Lipman, Barton L. 1999. "Decision Theory without Logical Omniscience: Toward an Axiomatic Framework for Bounded Rationality." *The Review of Economic Studies*, 66(2): 339–61.

Loewenstein, George and Drazen Prelec. 1992. "Anomalies in Intertemporal Choice: Evidence and an Interpretation." *The Quarterly Journal of Economics*, 107(2): 573–97.

Low, Bobbi S., R. D. Alexander, and K. M. Noonan. 1987. "Human Hips, Breasts, and Buttocks: Is Fat Deceptive?" *Ethology and Sociobiology*, 8(4): 249–57.

Luce, Duncan and Howard Raiffa. 1957. *Games and Decisions*. New York: Wiley.

Mazur, James E. 1984. "Tests of an Equivalence Rule for Fixed and Variable Reinforcer Delays." *Journal of Experimental Psychology: Animal Behavior Processes*, 10(4): 426–36.

Mazur, James E. 1986. "Fixed and Variable Ratios and Delays: Further Tests of an Equivalence Rule."

Journal of Experimental Psychology: Animal Behavior Processes, 12(2): 116–24.

Mazur, James E. 1987. "An Adjusting Procedure for Studying Delayed Reinforcement," in *Quantitative Analyses of Behaviour: The Effect of Delay and of Intervening Events on Reinforcement Value*. Michael L. Commons, James E. Mazur, John A. Nevin, and Howard Rachlin, eds. Hillsdale, NJ: Lawrence Erlbaum, 55–73.

McCabe, Kevin A., Stephen J. Rassenti, and Vernon L. Smith. 1998. "Reciprocity, Trust, and Payoff Privacy in Extensive Form Bargaining." *Games and Economic Behavior*, 24(1–2): 10–24.

McKelvey, Richard D. and Thomas R. Palfrey. 1992. "An Experimental Study of the Centipede Game." *Econometrica*, 60(4): 803–36.

McKelvey, Richard D. and Thomas R. Palfrey. 1995. "Quantal Response Equilibria for Normal Form Games." *Games and Economic Behavior*, 10(1): 6–38.

McKelvey, Richard D. and Thomas R. Palfrey. 1998. "Quantal Response Equilibria for Extensive Form Games." *Experimental Economics*, 1(1): 9–41.

Milgrom, Paul. 2004. *Putting Auction Theory to Work*. Cambridge: Cambridge University Press.

Milton, Katharine. 1988. "Foraging Behavior and the Evolution of Primate Cognition," in *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes and Humans*. Richard W. Byrne and Andrew Whiten, eds. Oxford: Clarendon Press: 285–305.

Møller, Anders. 1988. "Female Choice Selects for Male Sexual Tail Ornaments in the Monogamous Swallows." *Nature*, 332(6165): 640–42.

Nash, John F. 2002. "Noncooperative Games," in *The Essential John Nash*. Harold W. Kuhn and Sylvia Nasar, eds. Princeton: Princeton University Press: 85–98.

Ochs, Jack and Alvin E. Roth. 1989. "An Experimental Study of Sequential Bargaining." *American Economic Review*, 79(3): 355–84.

Pinker, Steven. 1997. *How the Mind Works*. New York: W. W. Norton.

Plott, Charles R. 1996. "Rational Individual Behavior in Markets and Social Choice Processes: The Discovered Preference Hypothesis," in *The Rational Foundations of Economic Behavior*. Kenneth J. Arrow, Enrico Colombatto, Mark Perlman and Christian Schmidt, eds. New York: St. Martin's Press, 225–50.

Plott, Charles R. and Vernon L. Smith. 1978. "An Experimental Examination of Two Exchange Institutions." *The Review of Economic Studies*, 45(1): 133–53.

Plott, Charles R. and Kathryn Zeiler. 2003 "The Willingness to Pay/Willingness to Accept Gap, the 'Endowment Effect,' Subject Misconceptions and Experimental Procedures for Eliciting Valuations," Mimeo. California Institute of Technology.

Postlewaite, Andrew. 1998. "The Social Basis of Interdependent Preferences." *European Economic Review*, 42(3–5): 779–800.

Prasnikar, Vesna and Alvin E. Roth. 1992. "Considerations of Fairness and Strategy: Experimental Data from Sequential Games." *The Quarterly Journal of Economics*, 107(3): 865–88.

Rabin, Matthew. 1993. "Incorporating Fairness into Game Theory and Economics." *American Economic Review*, 83(5): 1281–302.

Rabin, Matthew. 2000. "Risk Aversion and Expected-Utility Theory: A Calibration Theorem." *Econometrica*, 68(5): 1281–92.

Rabin, Matthew and Richard H. Thaler. 2001. "Anomalies: Risk Aversion." *Journal of Economic Perspectives*, 15(1): 219–32.

Rassenti, Stephen J., Vernon L. Smith, and Robert L. Bulfin. 1982. "A Combinatorial Auction Mechanism for Airport Time Slot Allocation." *Bell Journal of Economics*, 13(2): 402–17.

Ridley, Matt. 1993. *The Red Queen: Sex and the Evolution of Human Nature*. New York: Penguin Books.

Robson, Arthur J. 1992. "Status, the Distribution of Wealth, Private and Social Attitudes to Risk." *Econometrica*, 60(4): 837–57.

Robson, Arthur J. 1996a. "A Biological Basis for Expected and Non-expected Utility." *Journal of Economic Theory*, 68(2): 397–424.

Robson, Arthur J. 1996b. "The Evolution of Attitudes to Risk: Lottery Tickets and Relative Wealth." *Games and Economic Behavior*, 14(2): 190–207.

Robson, Arthur J. 2001. "The Biological Basis of Economic Behavior." *Journal of Economic Literature*, 39(1): 11–33.

Robson, Arthur J. 2001. "Why Would Nature Give Individuals Utility Functions?" *Journal of Political Economy*, 109(4): 900–14.

Ross, Sheldon. 1996. *Stochastic Processes*. New York: Wiley.

Roth, Alvin E. 1987. "Laboratory Experimentation in Economics," in *Advances in Economic Theory: Fifth World Congress of the Econometric Society*. Truman Bewley, ed. Cambridge: Cambridge University Press, 269–99.

Roth, Alvin E. 1988. "Laboratory Experimentation in Economics: A Methodological Overview." *Economic Journal*, 98(393): 974–1031.

Roth, Alvin E. 1991. "Game Theory as a Part of Empirical Economics." *Economic Journal*, 101(404): 107–14.

Roth, Alvin E. 1993. "The Early History of Experimental Economics." *Journal of the History of Economic Thought*, 15(2): 184–209.

Roth, Alvin E. 1994. "Let's Keep the Con out of Experimental Econ.: A Methodological Note." *Empirical Economics*, 19(2): 279–89.

Roth, Alvin E. 1995. "Bargaining Experiments," in *Handbook of Experimental Economics*. John H. Kagel and Alvin E. Roth, eds. Princeton: Princeton University Press, 253–348.

Roth, Alvin E. and Ido Erev. 1995. "Learning in Extensive-Form Games: Experimental Data and Simple Dynamic Models in the Intermediate Term." *Games and Economic Behavior*, 8(1): 164–212.

Roth, Alvin E. and Michael W. K. Malouf. 1979. "Game-Theoretic Models and the Role of Information in Bargaining." *Psychological Review*, 86(6): 574–94.

Roth, Alvin E. and J. Keith Murnighan. 1982. "The Role of Information in Bargaining: An Experimental Study." *Econometrica*, 50(5): 1123–42.

Roth, Alvin E., Vesna Prasnikar, Masahiro Okuno-Fujiwara, and Shmuel Zamir. 1991. "Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study." *American Economic Review*, 81(5): 1068–95.

Roth, Alvin E. and Francoise Schoumaker. 1983. "Expectations and Reputations in Bargaining: An Experimental Study." *American Economic Review*, 73(3): 362–72.

Rubinstein, Ariel. 1998. *Modeling Bounded Rationality*. Cambridge: MIT Press.

Rubinstein, Ariel. 2001. "Comments on the Risk and Time Preferences in Economics," Mimeo. Tel Aviv University.

Rubinstein, Ariel. 2003. "'Economics and Psychology'? The Case of Hyperbolic Discounting." *International Economic Review*, 44(4): 1207–16.

Samuelson, Larry. 2001. "Analogies, Adaptation, and Anomalies." *Journal of Economic Theory*, 97(2): 320–66.

Samuelson, Larry. 2004. "Information-Based Relative Consumption Effects." *Econometrica*, 72(1): 93–118.

Samuelson, Larry and Jeroen Swinkels. 2001. "Information and the Evolution of the Utility Function," SSRI Working Paper 2001-06, University of Wisconsin.

Sandroni, Alvaro. 2003. "The Reproducible Properties of Correct Forecasts." *International Journal of Game Theory*, 32(1): 151–59.

Savage, Leonard J. 1972. *The Foundations of Statistics*. New York: Dover Publications.

Segal, Uzi and Joel Sobel. 2003. "Tit for Tat: Foundations of Preferences for Reciprocity in Strategic Settings," Mimeo. Johns Hopkins University and University of California, San Diego.

Selten, Reinhard. 1965. "Spieltheoretische Behandlung eines Oligopolmodells mit Nachfrageträgheit." *Zeitschrift fur die Gesamte Staatswissenschaft*, 121: 301–24 and 667–89.

Selten, Reinhard. 1975. "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive-Form Games." *International Journal of Game Theory*, 4(1–2): 25–55.

Selten, Reinhard, Abdolkarim Sadrieh, and Klaus Abbink. 1999. "Money Does Not Induce Risk Neutral Behavior, but Binary Lotteries Do Even Worse." *Theory and Decision*, 46(3): 211–49.

Slonim, Robert and Alvin E. Roth. 1998. "Learning in High Stakes Ultimatum Games: An Experiment in the Slovak Republic." *Econometrica*, 66(3): 569–96.

Smith, Cedric A. B. 1961. "Consistency in Statistical Inference and Decision." *Journal of The Royal Statistical Society, Series B*, 23(1): 1–37.

Smith, Vernon L. 1962. "An Experimental Study of Competitive Market Behavior." *Journal of Political Economy*, 70(2): 111–37.

Smith, Vernon L. 1964. "Effect of Market Organization on Competitive Equilibrium." *Quarterly Journal of Economics*, 78(2): 181–201.

Smith, Vernon L. 1965. "Experimental Auction Markets and the Walrasian Hypothesis." *Journal of Political Economy*, 73(4): 387–93.

Smith, Vernon L. 1976. "Bidding and Auctioning Institutions: Experimental Results," in *Bidding and Auctioning for Procurement and Allocation*. Y. Amihud ed. New York: New York University Press, 43–64.

Smith, Vernon L. 1982. "Microeconomic Systems as an Experimental Science." *American Economic Review*, 72(5): 923–55.

Smith, Vernon L. 1991. *Papers in Experimental Economics*. Cambridge: Cambridge University Press.

Smith, Vernon L. 2000. *Bargaining and Market Behavior: Essays in Experimental Economics*. Cambridge: Cambridge University Press.

Smith, Vernon L. 2003. "Constructivist and Ecological Rationality in Economics." *American Economic Review*, 93(3): 465–508.

Sober, Elliott and David Sloan Wilson. 1998. *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Cambridge: Harvard University Press.

Sozou, Peter D. 1998. "On Hyperbolic Discounting and Uncertain Hazard Rates." *Proceedings of the Royal Society of London, Series B*, 265(1409): 2015–20.

Sunder, Shyam. 2004. "Markets as Artifacts: Aggregate Efficiency from Zero-Intelligence Traders," in *Models of a Man: Essays in Honor of Herbert A. Simon*. Mie Augier and James G. March, eds. Cambridge: MIT Press, 501–19.

Thaler, Richard H. 1988. "Anomalies: The Ultimatum Game." *Journal of Economic Perspectives*, 2(4): 195–206.

Thaler, Richard H. 1992. *The Winner's Curse*. Princeton: Princeton University Press.

Thaler, Richard H. 1994. *Quasi-Rational Economics*. New York: Russell Sage Foundation.

Tversky, Amos and Daniel Kahneman. 1982. "Judgment under Uncertainity: Heuristics and Biases," in *Judgment under Uncertainity: Heuristics and Biases*. Daniel Kahneman, Paul Slovic, and Amos Tversky, eds. Cambridge: Cambridge University Press, 3–22.

Tversky, Amos and Daniel Kahneman. 1983. "Extensional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment." *Psychological Review*, 90(4): 293–315.

Walker, James M., Vernon L. Smith, and James C. Cox. 1990. "Inducing Risk-Neutral Preferences: An Examination in a Controlled Market Environment." *Journal of Risk and Uncertainty*, 3(1): 5–24.

Wason, Peter. 1966. "Reasoning," in *New Horizons in Psychology*. B. M. Foss, ed. Harmondsworth: Penguin Books.

Weibull, Jörgen W. 2004. "Testing Game Theory," in *Advances in Understanding Strategic Behaviour: Game Theory, Experiments and Bounded Rationality: Essays in Honour of Werner Güth*. Steffen Huck, ed. Basingstoke: Palgrave.

Winter, Eyal and Shmuel Zamir. 1997. "An Experiment with Ultimatum Bargaining in a Changing Environment," Discussion Paper 159, The Hebrew University, Center for Rationality and Interactive Decision Theory.